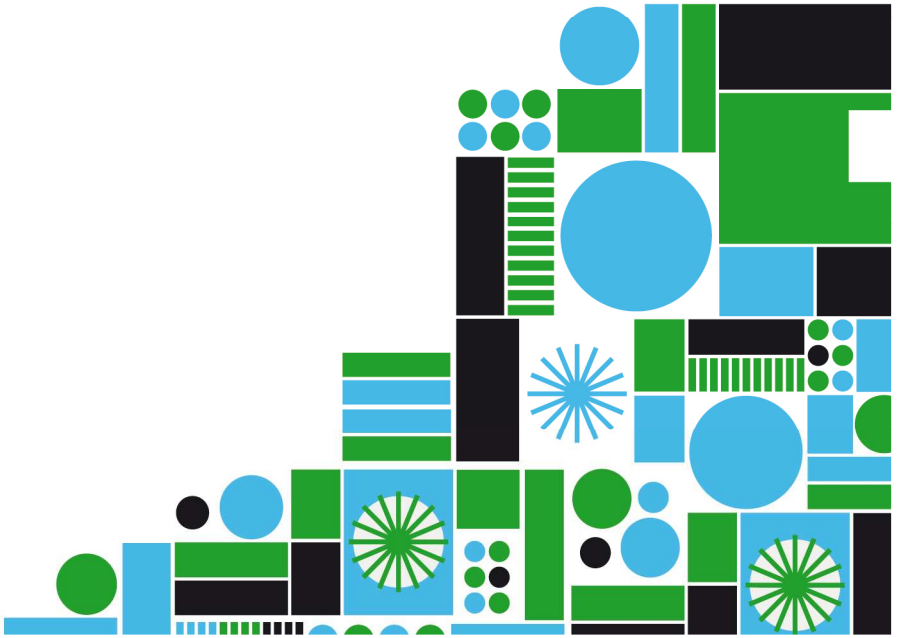
La Journée Technologique

13 janvier - Bois Colombes

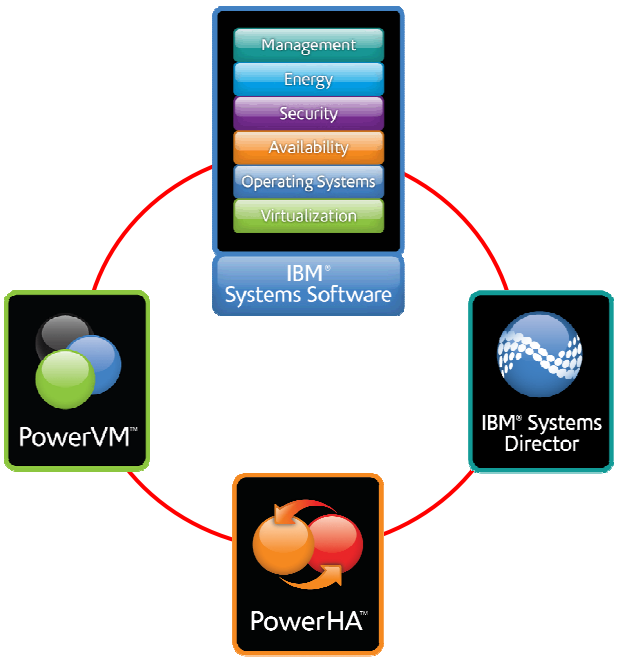
DATA CENTER : nouveaux enjeux, nouvelles solutions



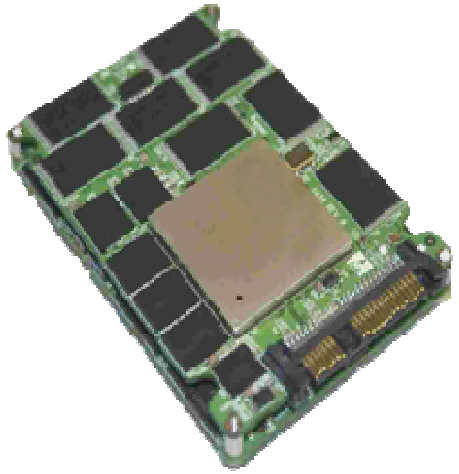
La Journée Technologique
13 janvier - Bois Colombes

La technologie Solid State Device

Didier FERAUD
FTSS Power Systems STG
didier_feraud@fr.ibm.com



Agenda

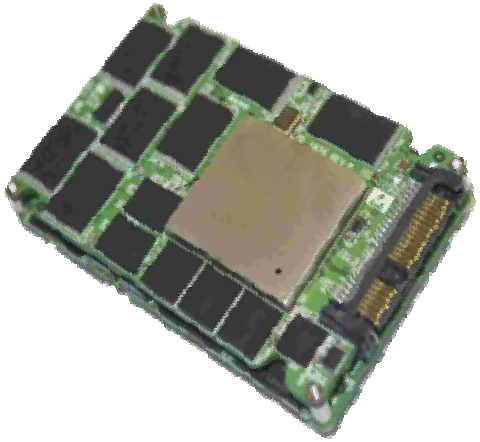


- Positionnement produit
- La technologie des SSD
- L'offre IBM
- IBM i et SSD
- Nouveautés Aout 2010
- Documentations

"Smarter Systems for a Smarter Planet"

© 2010 / 2011 IBM Corporation

Positionnement Produit



© 2010 / 2011 IBM Corporation

La technologie SSD: un coup de pouce pour les serveurs et le stockage !!!

Réduit les délais: temps de réponse des transactions, durée de traitement des batchs

- *réduction des temps d'attente des E/S (de 10 à 100 fois)*

Réduit les infrastructures de stockage: diminution des coûts liés aux contrôleurs et unités de disque, énergie, refroidissement, occupation au sol ….

- *augmentation des possibilités des E/S par élément de stockage (dans un rapport de 2 à 4 fois), réduction du nombre de bras d'accès nécessaire à la performance, …*

Réduit l'infrastructure serveur: capacité des mémoires DRAM, coût et énergie

Green!

- *pagination très rapide des OS ou de l'hyperviseur (réduction dans un rapport de 2)*

Améliore la disponibilité: augmentation du MTBF et meilleure anticipation des erreurs

- *amélioration du temps d'établissement des points de synchro et des dumps, moins de composants électromécaniques, reconstructions de contexte plus rapide…*

Améliore la gestion du changement: réduit les temps d'arrêt pour maintenance et changements pour les systèmes sensibles

- *augmentation du nombre d'E/S par seconde, réduction des délais d'IPL …*

Emergence de nouvelles possibilités: nouvelles fonctions et applications possibles

- *Amélioration des performances, des coûts ….*

© 2010 / 2011 IBM Corporation

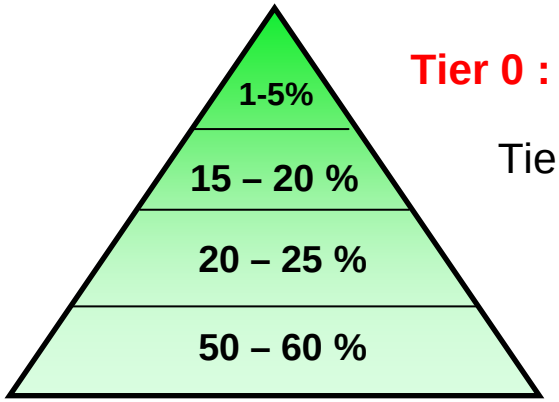
Définitions ….

✓ **C'est une unité de stockage constituée de mémoires flash ou mémoires persistantes qui peut être effacée électriquement et reprogrammée.**

✓ Inventée dans les années 80 par Toshiba, utilisée depuis les années 95 dans l'industrie militaire et aérospatiale en remplacement des disques traditionnels
Evolution des unités de stockage à base de DRAM (Dynamic Random Access Memory ou unité-RAM).

✓ Technologie présente dans les clefs USB flash et dans les cartes mémoire des caméscopes et autres appareils photos numériques (cartes micro SDHC…)

Classement des moyens de stockage en Tiers:



Tier 0 : Entreprise SSD (FC SSD, SAS SSD)

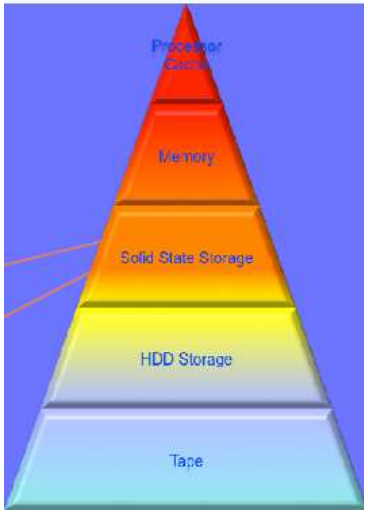
Tier 1 : HDD (FC HDD, SAS HDD)

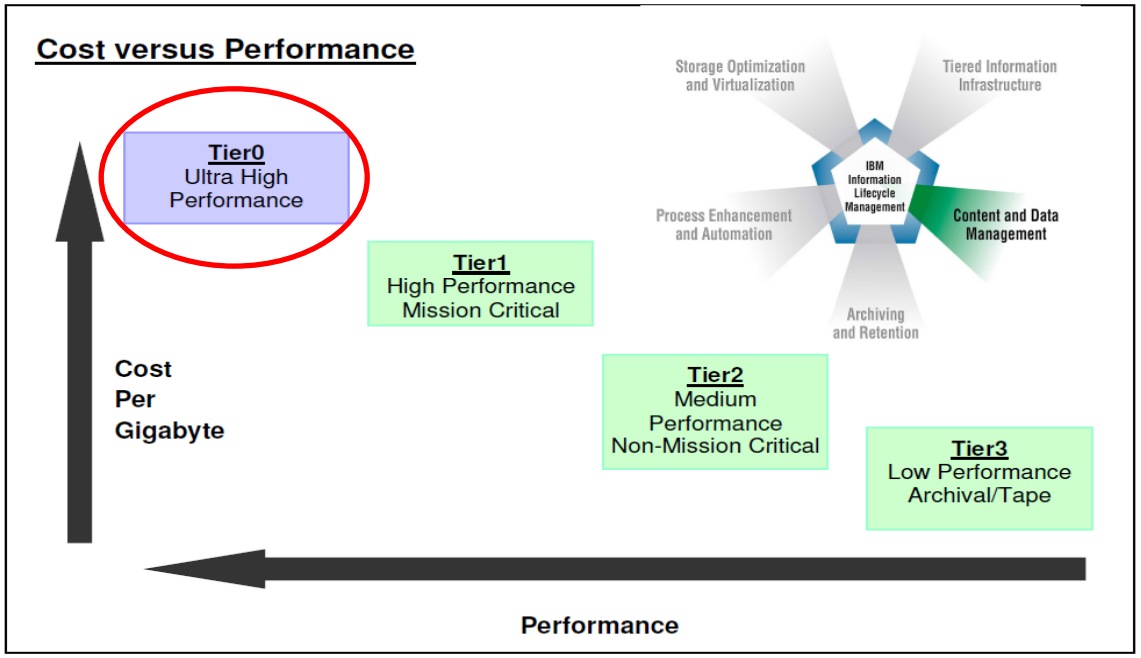
Tier 2 : SATA SSD & SATA HDD (interface Serial ATA)

Tier 3 : Bandes et stockage optique

© 2010 / 2011 IBM Corporation

Les différents niveaux de stockage des serveurs d'entreprise





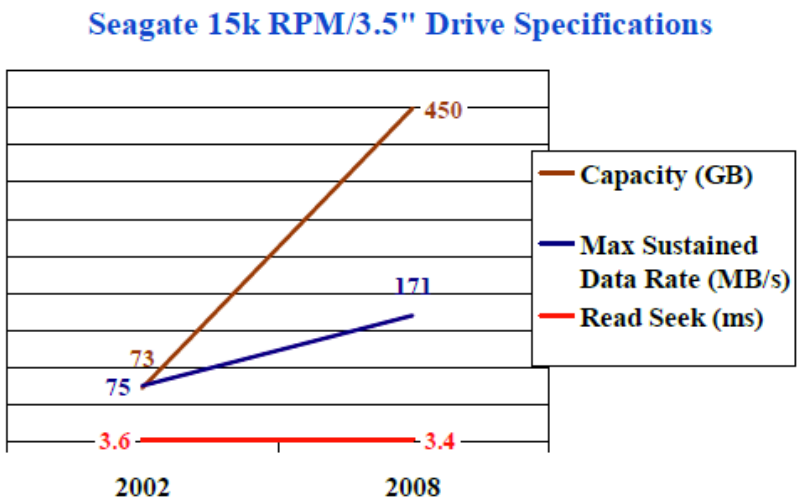
- L'emplacement des données permet de mettre en œuvre le tiering et d'optimiser la gestion des données au sein du système d'information.
- La combinaison de ces éléments constitue le socle d'une infrastructure de stockage dans le cadre de la mise en œuvre d'une infrastructure de PRA.

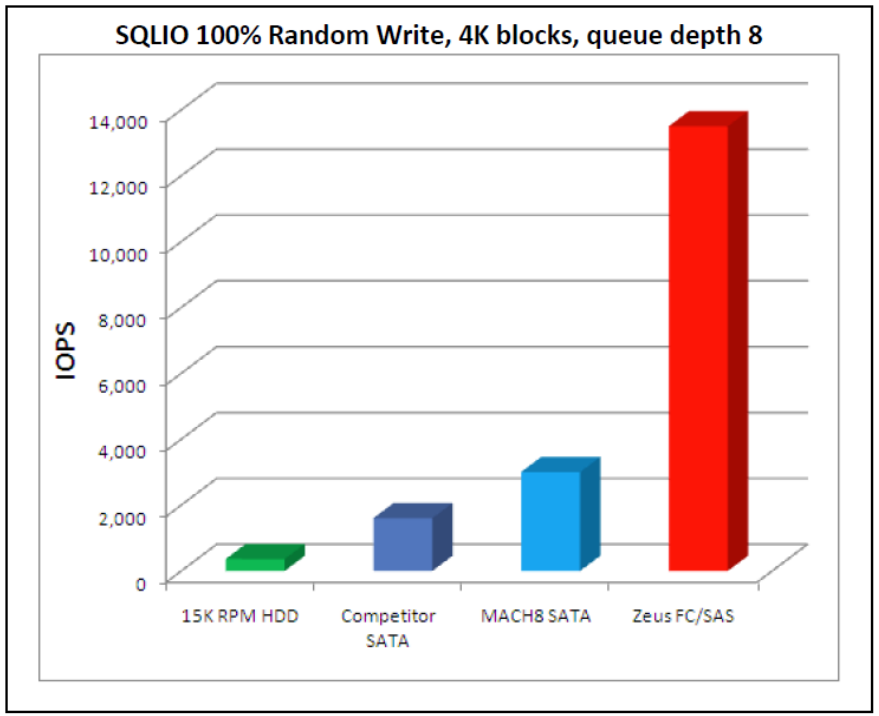
© 2010 / 2011 IBM Corporation

Nouveau métrique de référence

☞ Définition de nouveaux métriques pour positionner l'apport des SSD dans les serveurs d'entreprise… **prix par opération d'E/S par seconde**

(en HHD, le métrique de référence est le prix au Gigaoctet)





Source STEC Inc.

Attention à bien faire la distinction entre les SSD de classe Entreprise, beaucoup plus onéreux et les SSD de classe Grand Public, moins chers mais moins performants et moins robustes.

© 2010 / 2011 IBM Corporation

Performances des SSD en fonction des configurations

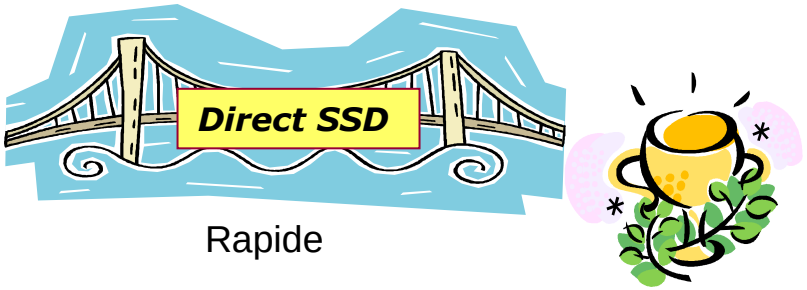
ns

Processors

Très, très, très, très, très rapide

Memory

Très, très, très rapide



Rapide

ms

Disk

Très, très lent comparativement

≤ 10 ns	~100 ns	~ 100.000 ns (100 µs)	900,000 -7,500,000 ns
Vitesse d'accès			

ns

Processors

Très, très, très, très, très rapide

Memory

Très, très, très rapide

Rapide

ms

Disk

Très, très lent comparativement

≤ 10 ns	~100 ns	~ 200.000 ns (200 µs)	1,000,000 - 8,000,000 ns
Mémoire vs Disque = x 8.000	**Vitesse d'accès**		***Proc vs Disque = x 80.000***

© 2010 / 2011 IBM Corporation

Performances des SSD

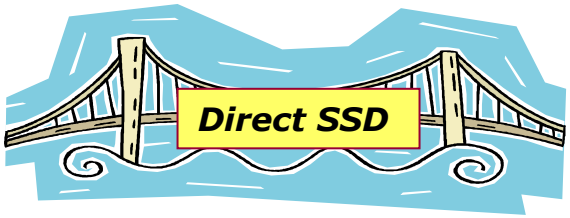
ns

Processors

Très, très, très, très, très rapide

Memory

Très, très, très rapide



Rapide

ms

Disk

Très, très lent comparativement

≤ 10 ns	~100 ns	~ 100.000 ns (100 µs)	900,000 -8,000,000 ns

Vitesse d'accès

Valeur médiane
~ 4.500.000 ns

~100 ms

~1 seconde	~17 minutes	~ 12 heures 1/2

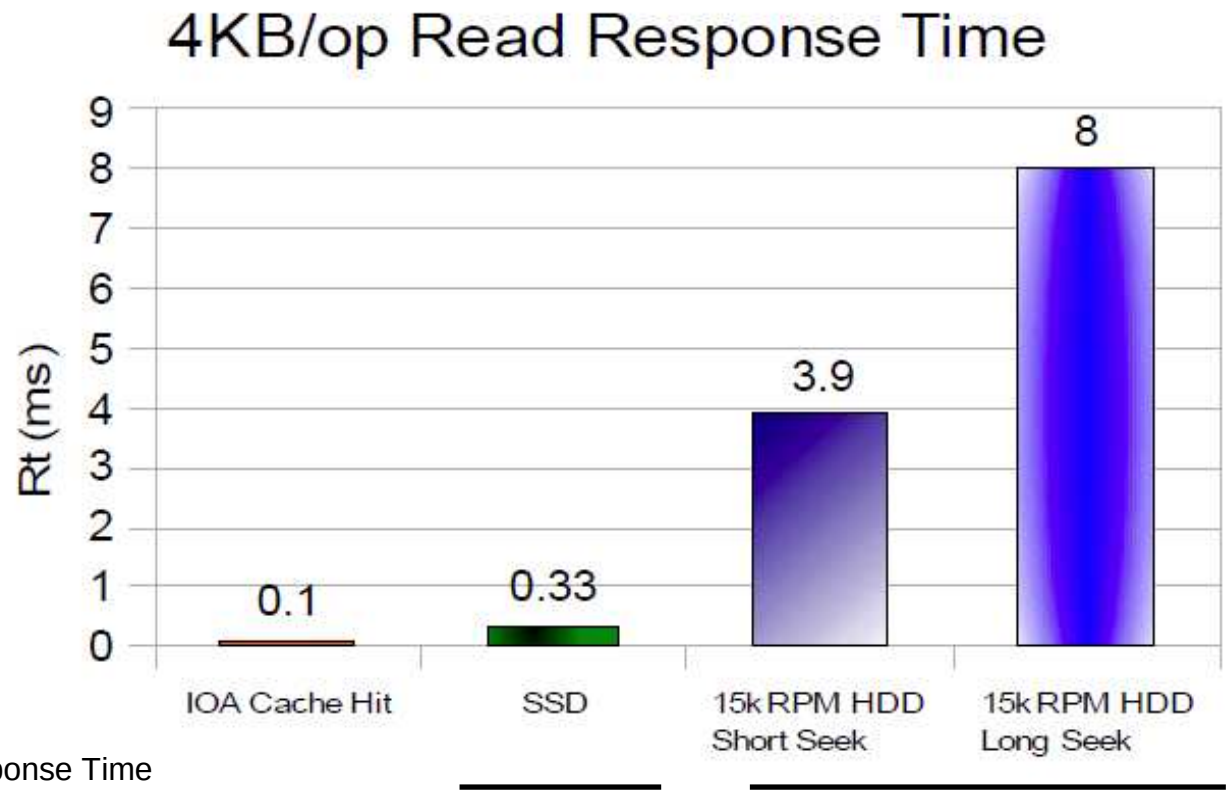
Contexte de temps homme

Mémoire vs Disque = x 8.000

Proc vs Disque = x 80.000

© 2010 / 2011 IBM Corporation

Temps de réponse des disques SSD et HDD en lecture



RT = Response Time

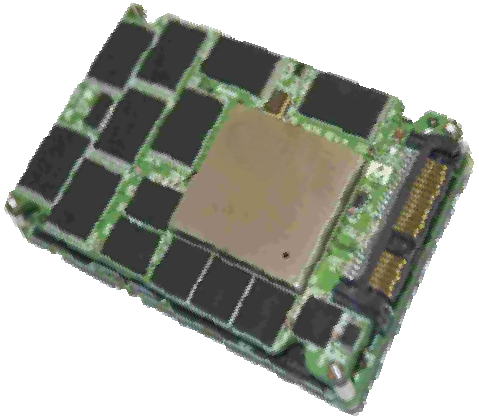
~ 28.000 op/s par unité

~ 150 à 200 op/s par bras

Le meilleur résultat est obtenu sur des travaux présentant un grand nombre de lecture aléatoires

© 2010 / 2011 IBM Corporation

Technologie des SSD



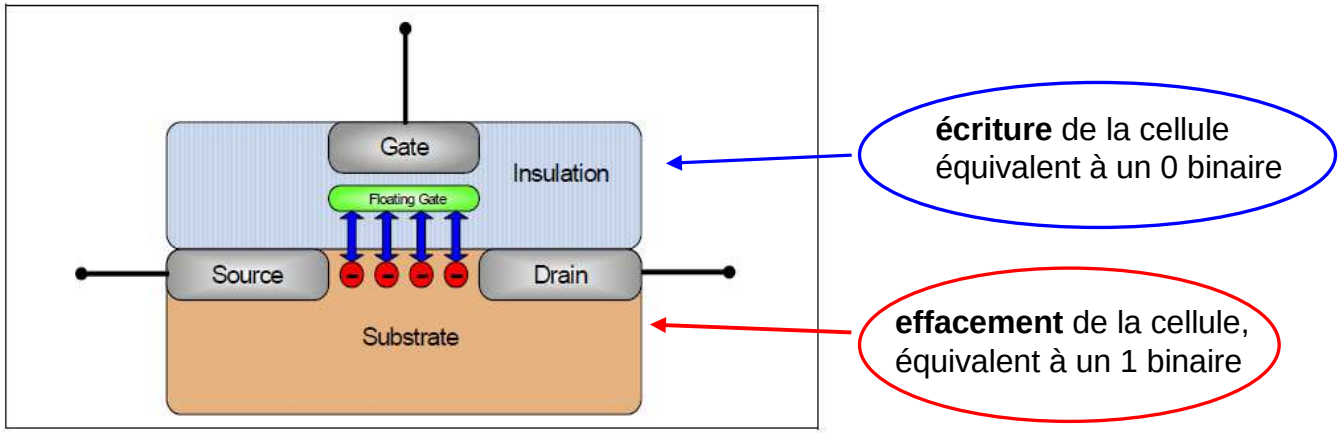
© 2010 / 2011 IBM Corporation

Les principes de la mémoire flash NAND

Une mémoire flash stocke l'information dans un groupe de cellules programmables. Chaque cellule élémentaire est constituée d'un transistor à effet de champs ou effet tunnel (effet *Fowler-Nordheim),* capable de déplacer des électrons en fonction des courants électriques appliqués au semi-conducteur.

Ecriture …Tension de 12v sur la grille de contrôle et tension de 7v entre drain et source

Lecture … mesure de la résistance de la grille flottante par courant de 5v entre grille de contrôle et une des deux électrodes.



Le processus de lecture est plus rapide que celui d'écriture ou d'effacement car on ne doit pas remplir ou vider la grille flottante avec des électrons, ce qui explique la différence de performance qui existe entre lecture et écriture.

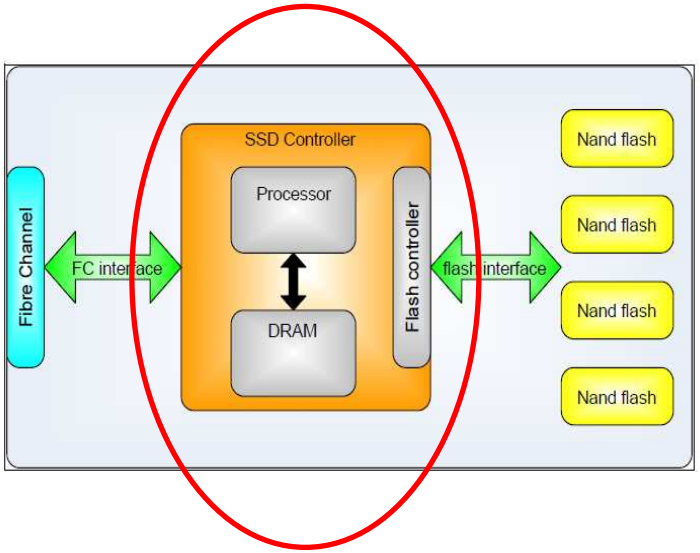
En technologie MLC (plus d'un bit par cellule), la quantité de courant qui circule est mesurée afin d'évaluer plus précisément le niveau de charge de la « floating gate » et déterminer la valeur des bits de la cellule.

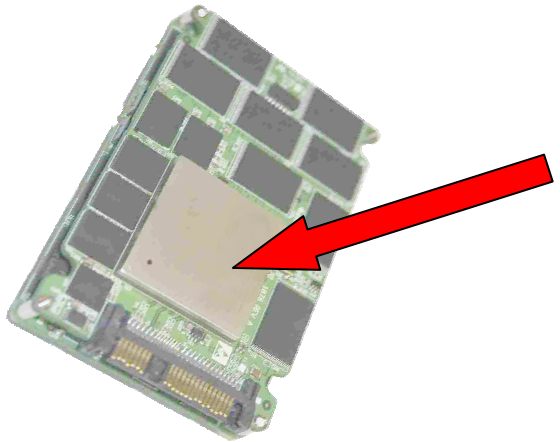
© 2010 / 2011 IBM Corporation

La lecture des données dans une mémoire flash NAND

La mémoire flash NAND travaille avec un bus série, en accès séquentiel. Il est impossible d'accéder directement à un bit en particulier. Pour accéder à une information précise, on doit charger une partie des données (ou page) dans une petite mémoire DRAM assimilable à un cache et ensuite lire ce que l'on souhaite dans cette mémoire.

Cette mémoire contient aussi un répertoire de placement des blocs et gère l'équilibrage dynamique de l'utilisation des cellules *(Static and Dynamic Wear-Leveling)*





Le contrôleur s'occupe, entre autres choses, de gérer la copie de la page dans la mémoire DRAM *(shadowing)* et de la gestion des blocs et de leur intégrité.

© 2010 / 2011 IBM Corporation

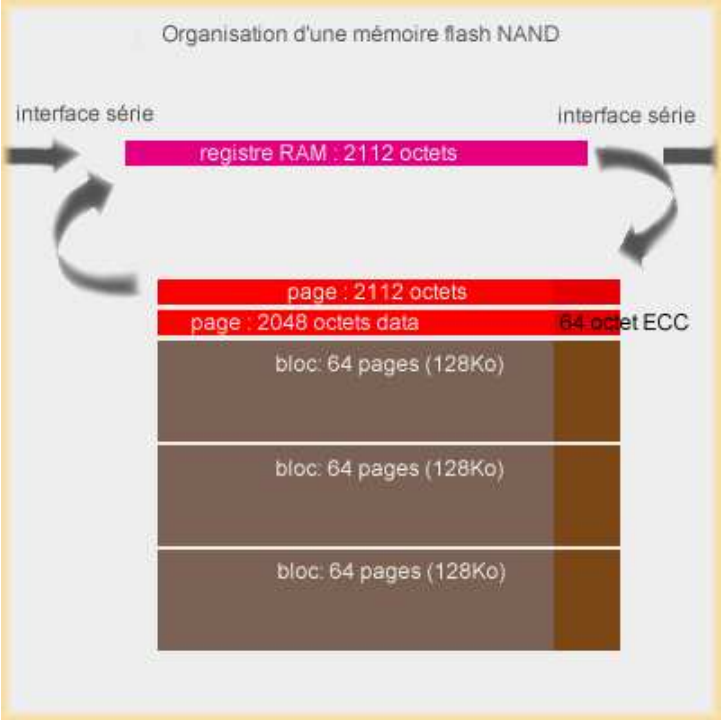
La mémoire flash NAND et l'organisation en blocs

☞ Lecture sur une page, écriture sur un bloc

- la page (2048 octets de donnée + 64 octets ECC) est l'unité minimale pour la lecture
- écriture au niveau du bloc (64 pages ou 128 Ko):

La moindre écriture oblige l'effacement du bloc de données complet (128 Ko) avant l'écriture (ou programmation) d'une nouvelle valeur; d'où la différence de performance entre lecture et écriture.

En pratique, la lecture et l'écriture de petits fichiers (< à 128 Ko) sont peu performantes



Lecture ….

Accès à une page de données : ~ 25 µs

(temps de copie de la page dans la DRAM interne)

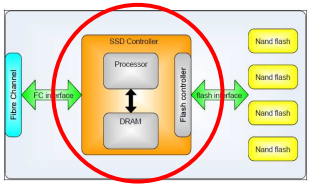
Accès aux autres pages résidentes du bloc : ~ 0.03 µs

Ecriture (ou programmation)…

Effacement / écriture d'un bloc de 64 pages : ~ 2 ms

© 2010 / 2011 IBM Corporation

Le contrôleur interne



C'est la partie la plus importante de l'unité. Puce de type SoC *(System on a Chip),* il va gérer l'accès aux données mais aussi assurer la gestion de l'ensemble des cellules NAND

Les principales fonctions du contrôleur utilsé par IBM (autres que le transfert des données):

- ✓ **Algorithme de « Wear-Leveling »**
  - Dynamic Wear-Leveling
    distribution aléatoire des écritures sur les blocs libres ou effacés (usure équilibrée)
  - Static Wear-Leveling
    déplacement des données statiques en fonction de la fréquence d'utilisation des blocs
- ✓ **Algorithme du « Bad block »**

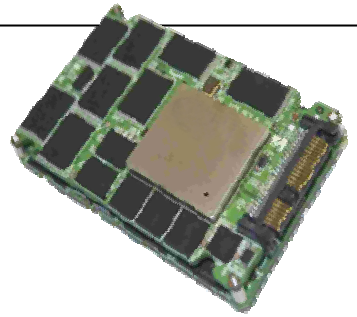
  détection des blocs défectueux lors de leur utilisation, en général dans le cycle effacement / programation et exclusion / remplacement dans la rotation des blocs
- ✓ **Algorithme de l' « Over Provisioning »**

  gestion d'une partie de la mémoire du SSD réservée et inaccessible à l'utilisateur (~75%) en combinaison avec la fonction « static wear-leveling »
- ✓ **Algorithme « Error Correction Code / Error Detection Code » *(ECC/EDC)***

  maintient l'intégrité de la donnée stockée en relisant la dernière donnée écrite avec déplacement /retrait complet du bloc si elle est altérée (EDC) et correction de la donnée en cours de lecture ou écriture avec gestion du bloc endommagé (ECC)

© 2010 / 2011 IBM Corporation

Mémoire SLC et mémoire MLC



Il existe deux types de mémoire NAND:

- la mémoire **SLC**, ou *Single Layer Cell* qui stocke un seul bit dans la grille flottante
  - *supporte en moyenne 100.000 cycles d'écriture*
- la mémoire **MLC** , ou *Multi Layer Cell* qui stocke plusieurs bits dans la même cellule
  - *on double la capacité de stockage en gardant la même taille physique*
  - *80% du débit d'une SLC en lecture, 50 % des performances en écriture*
  - *supporte en moyenne 10.000 cycles d'écriture*

Les données stockées dans ce type de mémoire persistent pendant une dizaine d'années sans être alimentée électriquement.

© 2010 / 2011 IBM Corporation

La durée de vie d'une mémoire flash

Une des particularités de la mémoire flash est sa durée de vie relativement limitée.

Les deux principales causes de ce phénomène ..

✓ l'oxyde utilisé pour séparer les grilles

Par construction, les électrons traversent cet oxyde en fonction des opérations d'écritures / effacements

De temps en temps, des électrons peuvent rester captifs de cet oxyde et perturber les lectures / écritures en provoquant un état indéterminé de la cellule

✓ la structure de la grille flottante et les tensions appliquées

Les différentes tensions appliquées à la grille (~12 volts en saturation pour écrire un 0 binaire) et 5 volts pour lire la cellule peuvent endommager la cellule avec le temps et la rendre inutilisable

Technologie NAND	Nombre d'écritures d'une cellule *
SLC	100.00 cycles
MLC	10.000 cycles

* Valeurs moyennes couramment admises

© 2010 / 2011 IBM Corporation

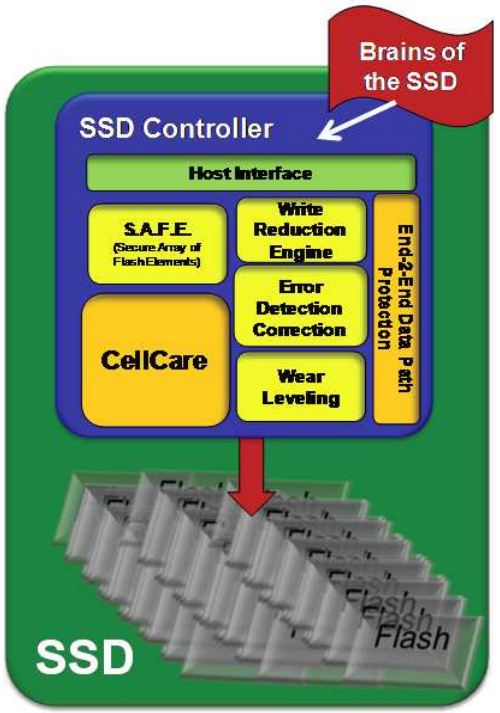
Utilisation de la technologie eMLC

Performance, fiabilité, disponibilité et robustesse des SSD de classe entreprise sont fortement dépendantes de la conception des contrôleurs des SSD.

La dernière génération des micro contrôleurs STEC pour SSD intègre la technologie CellCare™ , améliorant les performances des mémoires flash NAND de type MLC grâce à la gestion dynamique des accès à la mémoire, à l'utilisation de nouvelles techniques de traitement avancé des signaux optimisant la gestion de l'usure, aux nouveaux algorithmes de gestion des données et à l'amélioration du code de correction d'erreurs (ECC).

De plus, l'utilisation de la technologie S.A.F.E. (*Secure Array of Flash Element)* élimine virtuellement les problèmes d'intégrité des données (dysfonctionnement et corruption/perte de données).
Cette combinaison de technologies appliquée aux mémoires MLC réduit drastiquement les incidents au niveau des composants et fournit la fiabilité des données requise par les applications d'entreprise et les serveurs de stockage.

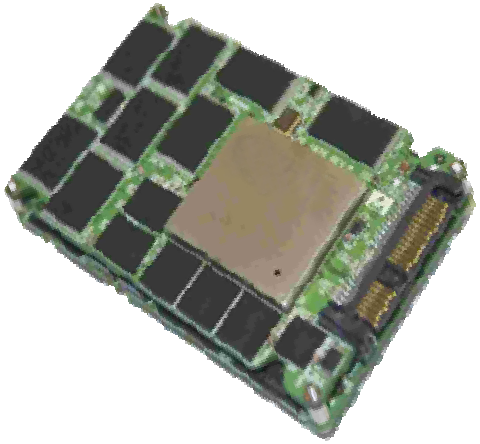
IBM intègre dès à présent les versions SAS 6Gbps et Fiber Channel 4Gbps des unités SSD ZeusIOPS de STEC basées sur la technologie eMLC dans sa gamme DS8800 et DS8700, ainsi que dans la nouvelle gamme des baies de stockage intermédiaires Storwize V7000.



Contrôleur SSD STEC avec fonction CellCare

© 2010 / 2011 IBM Corporation

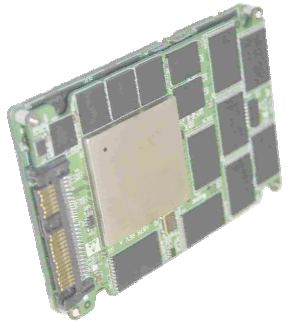
L’offre IBM



© 2010 / 2011 IBM Corporation

L'offre SSD sur les serveurs Power

Produit développé en collaboration avec la société STEC



Capacité utile : 69 Go (capacité réelle 128 GO, soit 83% de réserve)

Utilisation de la technologie mémoire flash SLC *(Single Level Cell)*

… un bit stocké par cellule (plus de 100.000 cycles d'écritures)

Supportés dans les modèles :

- *Power 520 et 550 / 750 / 755 HPC*
- *Power 560, 570, 770 et 780*
- *Power 595*
- Tiroir disques EXP 12S
- Tiroir PCIe 12X #5877 *(18 slots disque SFF)*

Interface SAS à 3 Gigabit/s

- *Inserts aux formats 2,5 et 3,5 pouces*
- *Deux options : format AIX / Linux ou format IBM i*

Performance (débit) :

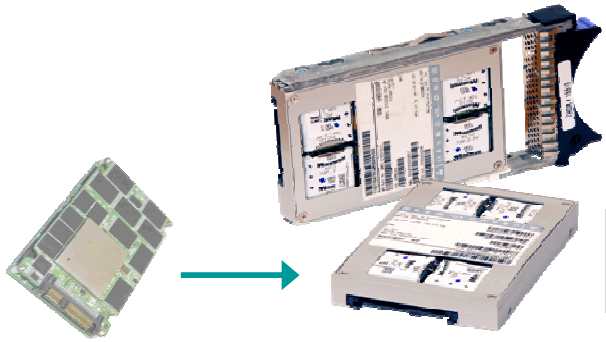
- **up to 240 MO/s (en lecture séquentielle)**
- **up to 125 MO/s (en écriture séquentielle)**

Nombre d'opérations par seconde en accès aléatoire

- **28.000 E/S par seconde**

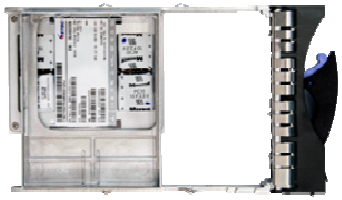
Temps d'accès moyen en lecture :

- **de 100 à 200 micro-secondes**



AIX #1890
IBM i #3586

Insert au format 2,5 pouces (SFF)



AIX #1909
IBM i #3587

Insert au format 3,5 pouces

Consommation électrique inférieure à la moitié de celle des disques 15 Kt/mn

© 2010 / 2011 IBM Corporation

Comparaison de prix ….

Prix tarif d'un disque SSD 69 GO de type EFD* = 7.002 € HT (Power 8204)

Prix tarif d'un disque SAS HDD 140 GO = 731 € HT (Power 8204)

Il faut donc réserver les disques SSD à l'usage qui aura le plus d'impact en terme de performances ou de rapport prix / performance

Utiliser les disques SSD uniquement lorsque le critère performances est primordial….

- *Utiliser les disques HDD lorsque le coût par GO est le critère de choix principal*
- *Utiliser une combinaison de disques SSD et HDD dans la même configuration sera très souvent une excellente solution*

Disques SSD

- Coût beaucoup plus élevé pour la même capacité
- Bien meilleure performance
- Moins d'unités = moins d'énergie et d'espace

Disques HDD

- Coût au GO le plus faible
- Peut imposer un nombre beaucoup plus important de disques pour avoir de bonnes performances

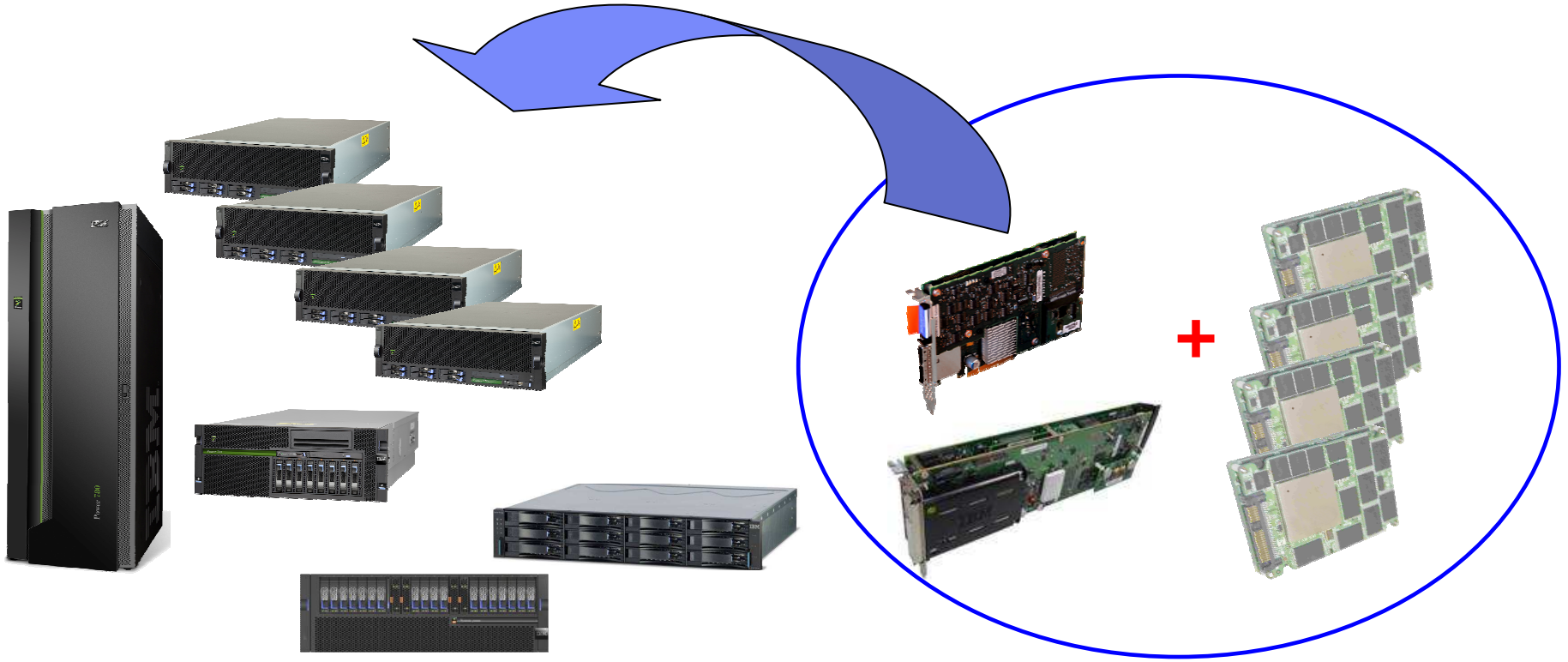
(*) EFD= Entreprise Flash Drive

Nota : Prix donnés à titre indicatif au 15/05/10, seul le tarif fait foi

© 2010 / 2011 IBM Corporation

Règles de configuration en système Power

- Afin d'optimiser la performance en écriture et garantir un niveau de protection minimum, il est impératif d'associer les unités SSD à des cartes contrôleur Raid5 afin de bénéficier des caches en écriture et des fonctions de protection intégrées dans ces cartes (#5903 en paire, #5904, #5908, …)
- Limiter le nombre d'unités SSD sur les contrôleurs (de 4 à 8) afin d'éviter la saturation de la carte contrôleur



© 2010 / 2011 IBM Corporation

Les caractéristiques des SSD en bref

- Les SSD s'installent …
  - 3.5-inch SAS drive bay of #5886 EXP 12S SAS Disk Drawer
  - 3.5-inch SAS drive bay of Power 560/570 CEC
  - SFF drive bay of Power 520/550 CEC 2.5-inch
- Les SSD supportent la fonction de "Boot" *(Load Source)*
- Les SSD supportent la correction de type ECC
- Les SSD ont une fonction de gestion de blocs en erreur

- Les SSD de type SFF & 3.5" sont pilotés par le contrôleur SAS comme un disque SAS classique
  - **Utilisation recommandée de contrôleur PCI-X 1.5 GB "write cache" #5904/#5908 ou double contrôleur #5903**
  - Fonctionnent aussi avec le contrôleur interne des Power 520/550 et Power 560/570
    - Solution moins performante que le contrôleur 1.5 GB, mais bien moins cher !!

- Le contenu des disques est protégé par les fonctions classiques des unités de disque SAS
  - Mirroring, RAID-5 ou RAID-6 (ainsi que la fonction "Hot Spare")

- Support des unités SSD à partir de l'i5 OS 5.4 et plus. La différence entre version se fera sur la gestion du placement des données plus ou moins automatisée *(fonction « i-balancer »)*

- **Règles de configuration:**
  - Contrôleur interne …. Mélange des unités SAS et SSD supporté
  - Fonction RAID interne …. Pas de mélange d'unités SAS et SSD
  - "Split Backplane" …. Pas de mélange d'unités SAS et SSD dans le même groupe d'unités
  - Tiroir PCIe 12X #5877 …. « Split » possible en deux groupes de 9 emplacements
  - Tiroir SAS EXP 12S: …. Supporte 8 unités SSD / Pas de mélange d'unités
    - ☛ *quatre slots restants non utilisés*

© 2010 / 2011 IBM Corporation

SSD et stockage externe

DS8700/8800

DS8700 .. SSD 73Go (FC/LFF), 146Go(FC/LFF), 600Go (FC/LFF)

Protection en Raid 5, 128 SSD maxi, 16 par paire de DA

Release 5.1 pour support « Easy Tiering »

DS8800 … SSD eMLC 300Go (SAS/SFF)

Release 6.0 et 6.1 (1H2011)

• *Utilisation de cartes #5735 de préférence et technologie P6+ ou P7*

Storwize V7000

- Unités SSD eMLC 300 Go (SAS/SFF)
- Fonction « Easy Tiering » de base
- Interface graphique de gestion

© 2010 / 2011 IBM Corporation

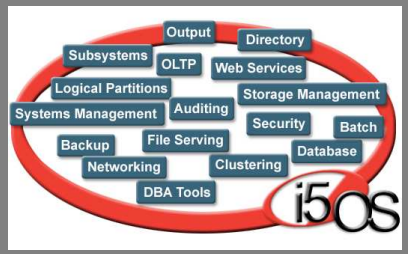
IBM i et SSD

© 2010 / 2011 IBM Corporation

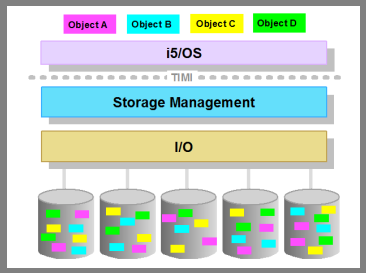
IBM i : les 5 points clés d'une architecture innovante

Intégration

IBM i comprend la base de données, la sécurité, les communications, la sauvegarde, etc.., le tout intégré et testé par IBM, et inclus dans le prix de la licence



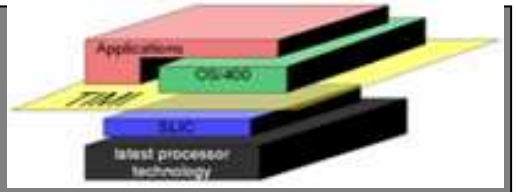
Gestion du stockage



IBM i répartit les données sur tous les disques, améliorant les performances d'accès tout en réduisant la gestion inhérente au placement de ces données

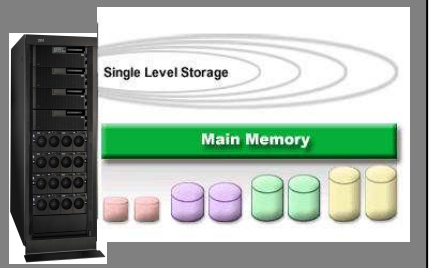
Technology Independent Machine Interface (TIMI)

Les nouvelles technologies matériel peuvent être implémentées sans affecter les applications

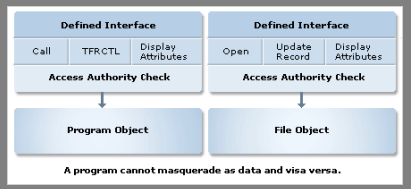


Espace adressable unique

IBM i considère mémoire & stockage comme une entité unique, automatisant et optimisant la gestion de l'ensemble



Architecture orientée objet



Chaque objet sur le système intègre l'ensemble des règles liées à son utilisation, améliorant son intégrité, sa sécurité et sa résistance aux virus

© 2010 / 2011 IBM Corporation

IBM i sur Power Systems … un délicat équilibre

POWER7 BUILT ON Power™

POWER6™ BUILT ON Power™

PROCESSEUR

POWER6+

POWER7

Management
Energy
Security
Availability
OS and Integration
Virtualization
Power™ Systems Software

DDR2

DDR3

MEMOIRE

Cartes PCIe

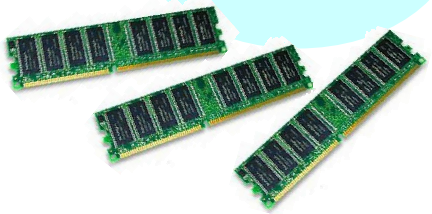
Boucles 12X DDR

Interface SAS

SFF 2.5inch

DISQUES

Batch !!!



PCI-X et PCIe

SSD !!

HDD

15 K rpm

© 2010 / 2011 IBM Corporation

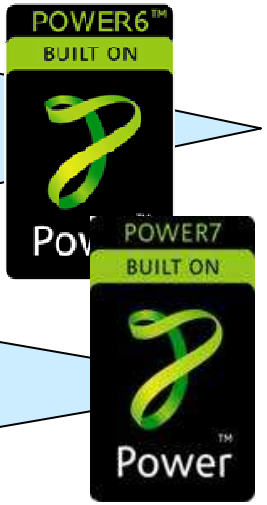
Transitions technologiques majeures des Entrées/Sorties

- **1. Interface SCSI vers SAS**

Disques SAS et SSD = 3.5-inch & SFF
Supports amovibles SAS & SATA

2. Adaptateurs PCI / PCI-X / PCI-X DDR vers PCIe

2008: PCIe disponible sur CEC 520/550/570 …, 2009: extension aux tiroirs d'E/S



3. Boucles RIO/HSL vers 12X

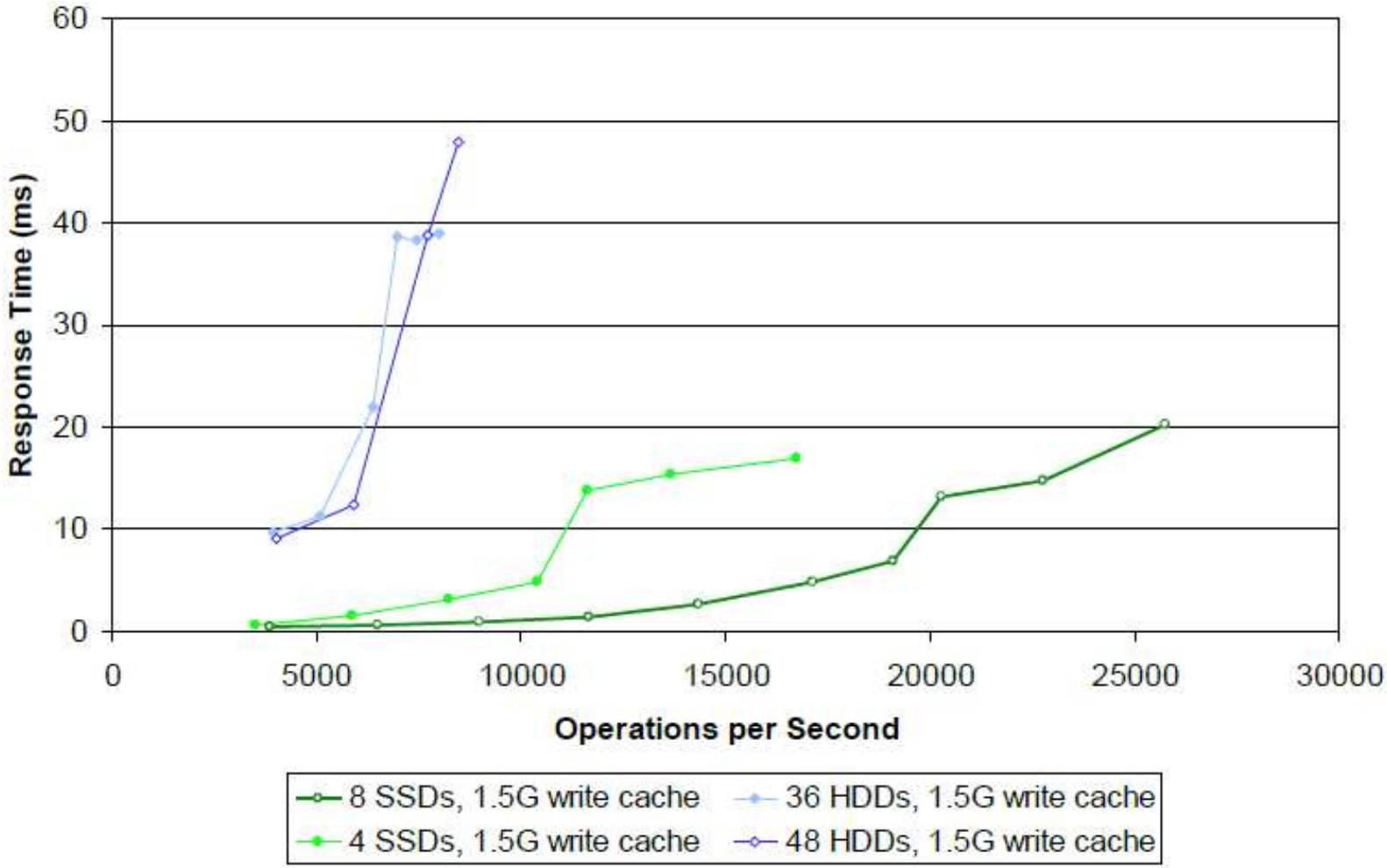
12X (SDR) et 12X DDR

4. IBM i : adaptateurs "IOP-based" vers "Smart IOA"

Plus d'IOP sur POWER 7

© 2010 IBM Corporation

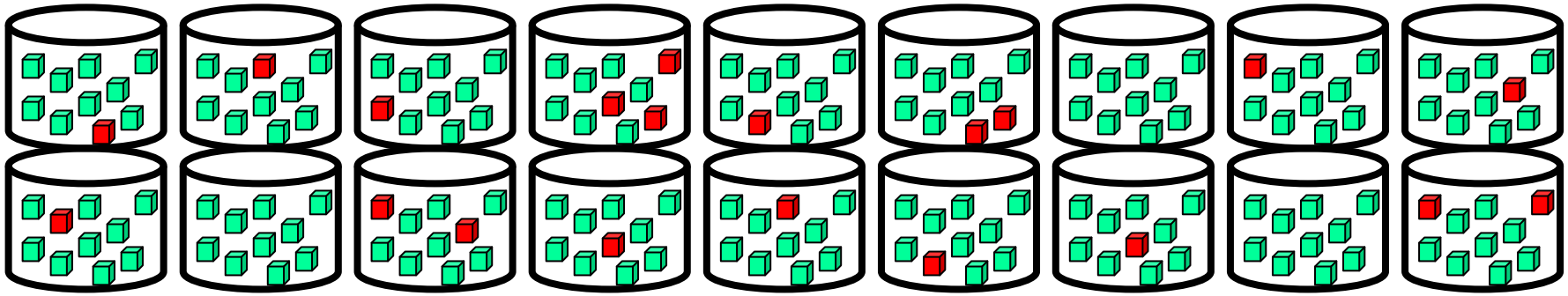
Comparaison SSD et HDD en environnement IBM i



© 2010 / 2011 IBM Corporation

Mélange SSD + HDD : une solution qui a du sens

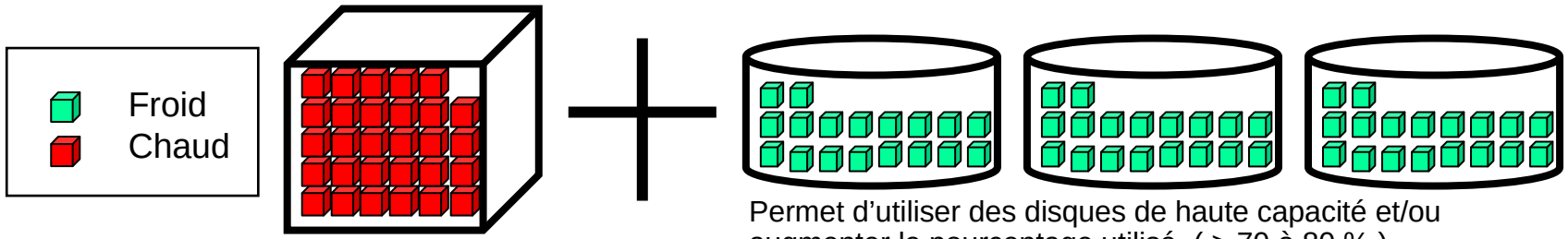
Par expérience, les bases de données ont un grand pourcentage de données qui sont utilisées à l'occasion ("froide") et un petit pourcentage de données utilisées fréquemment ("chaude")



Les données chaudes peuvent représenter 10 à 20 % de la capacité mais 80 à 90 % de l'activité

La solution SSD offre le meilleur rapport prix/performance quand elle est axée sur les données "chaudes"

La solution HDD offre le meilleur coût de stockage, axé sur la donnée "froide"…. approche de type HSM



SSD: proche de 100% de sa capacité

Permet d'utiliser des disques de haute capacité et/ou augmenter le pourcentage utilisé (> 70 à 80 %)

© 2010 / 2011 IBM Corporation

Positionnement des données “chaudes” et “froides”

AIX

Possibilité d’identifier et de positionner les fichiers “chauds”sur les SSD

- Les outils de performance clefs en AIX sont les utilitaires filemon & iostat.
- Les partenaires “Base de Données” fournissent des outils d’analyse de données “chaudes” (par exemple DB2 Snapshot).

Migration des données “chaudes”

- Outil Migratepv
- “Softek Migration Tool” fournit une suite d’outils et peut se combiner avec des prestations de service IBM
- Voir Total Storage Productivity Center

Un “white paper”sur SSD disponible vous donne de plus amples informations

Outils en cours de développement

IBM i

#1 La meilleure fonction automatique intégrée disponible dans l’industrie IT aujourd’hui

- La fonction “Trace and Balance” fait partie intégrante de l’IBM i
- Monitoring par partition ou ASP (Aux Storage Pool) pour déterminer les données chaudes ou froides
- A la demande, déplacement automatique des données chaudes vers les SSD et froides vers les HDD
- Processus à initier régulièrement

Les fichiers peuvent être placés automatiquement sur les SSD *(RSTLIB, RSTOBJ …)*

Certains objets spécifiques de type base de données peuvent être placés sur les SSD

Améliorations complémentaires de placement automatique intégrées dans la version IBM i 7.1

© 2010 / 2011 IBM Corporation

IBM i “Load Balancer”

Fonction totalement intégrée au système d’exploitation

Utilisation de la fonction “trace” pour collecter des données sur l’utilisation des partitions et ASP

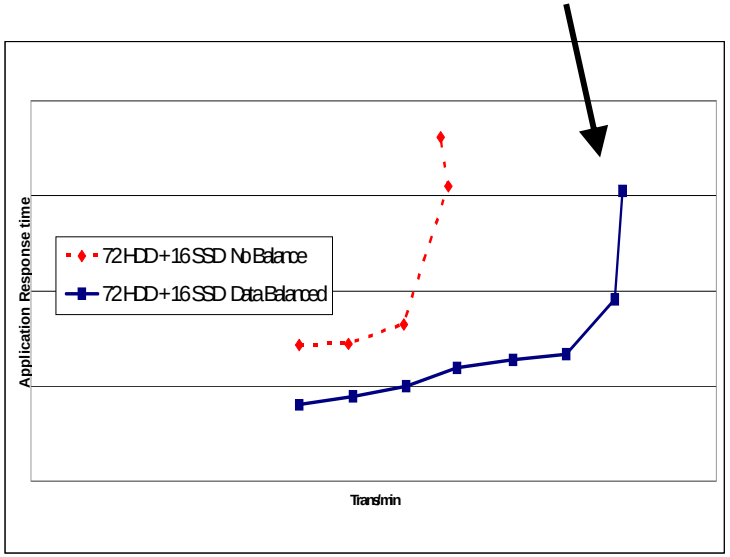
Activer la fonction trace pendant les périodes de pointe
- Désactiver cette fonction après une période de temps significative
- Impact négligeable sur la performance de la machine
- Utilisation des outils de performance pour identifier les données “chaudes”

Commande permettant automatiquement de déplacer les données « chaudes » vers les SSD et les données « froides » vers les HDD
- Peu d’impact en terme de performance, processus en arrière plan

Possibilité de “re-monitorer” et de “re-balancer” à n’importe quel moment
- Planification hebdomadaire ou mensuelle
- Probablement moins souvent si les données sont plutôt statiques

La fonction de placement des données chaudes/froides incluse dans le microcode de l’IBM i est plus performante que la distribution classique des données sur tous les disques du stockage.

Configuration: 72 HDD + 16 SSD



Si on veut analyser le % de données considérées comme “chaudes afin de déterminer le nombre de SSD nécessaire, il faut:
- Utiliser les résultats de l’outil PEX
- Interpréter les informations obtenues dans l’outil Performance Tools/400
- Utiliser l’outil (*) disponible sur le site IBM pour estimer l’intérêt d’implanter quelques disques SSD sur le serveur Power

(**) Download http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS3780

© 2010 / 2011 IBM Corporation

Amélioration du gestionnaire de stockage en IBM i pour les SSD

IBM i supporte la fonction de gestion de stockage hiérarchique

- Assure la collecte automatique des données de performance des E/S et gère le déplacement des données les plus sollicitées vers les “Solid State Drives” (SSD)
- ➢ Permet d’optimiser les investissements liés aux SSD à l’aide des amélioration de la gestion du stockage

DB2 for i supporte les SSD comme unités de prédilection

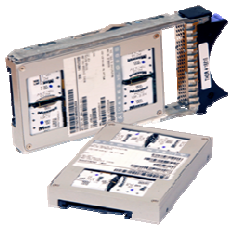
- Statistiques sur les lectures aléatoires DB2

Améliorations complémentaires pour la gestion des SSD

- Nouveaux utilitaires orientés SSD
- Possibilité de placer des objets de type IFS sur SSD
- Amélioration des instruments de performance
- Possibilité d’indiquer qu’une table à usage intensif ou un index soit spécifiquement positionné sur SSD:
  - **CRTPF** department **SRCFILE**(mjasrc/dds) **UNIT**(*SSD)
  - **CRTLF** departmentl **SRCFILE**(mjasrc/dds) **UNIT**(*SSD)
  - **CREATE TABLE** employee (c1 INT) **UNIT** SSD
  - **CREATE INDEX** employeeix **ON** mjatst.t2 (c1) **UNIT** SSD

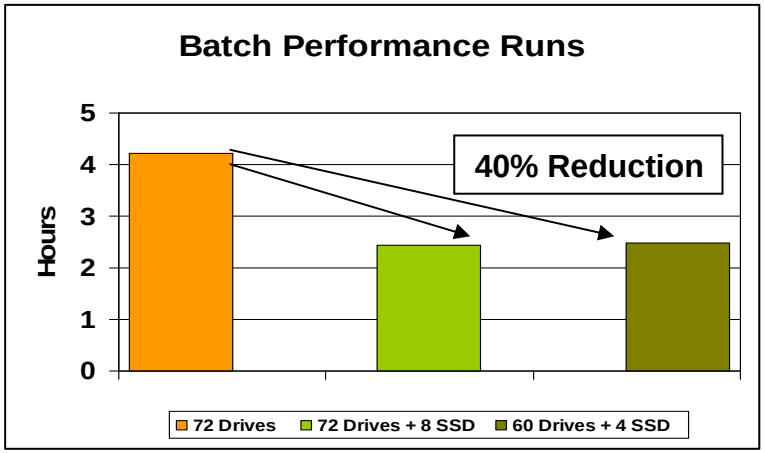
« SSD Analyzer Tool »

- Utilitaire destiné à déterminer si l’utilisation de SSD permet une amélioration des performances applicatives
- Compatible i5/OS 5.4, IBM i 6.1 et 7.1 (**)



V7R1

“Associated Bank” réduit la durée de son Batch de 40% avec les SSD (*)



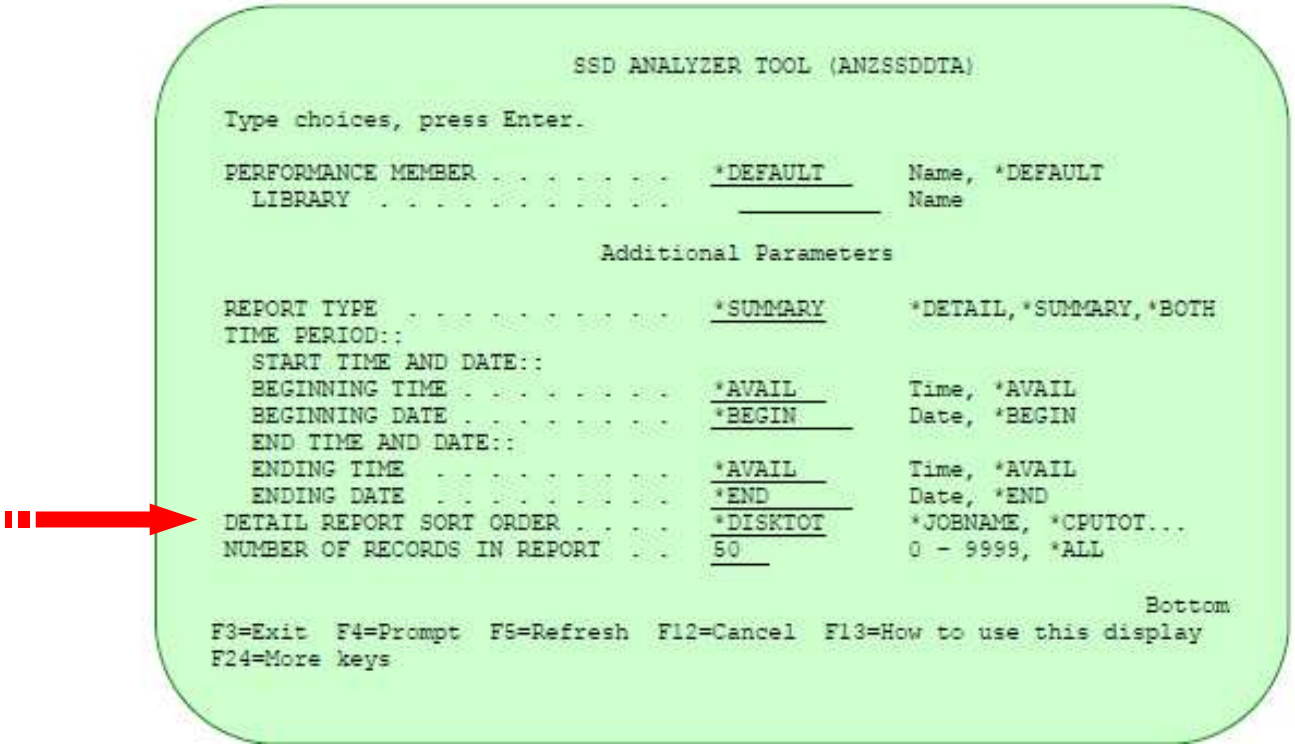
Les unités SSD peuvent améliorer les performances des longs traitements par lots et des requêtes tout en optimisant automatiquement le placement des données.

(*) http://www.ibmsystemsmagpowersystemsibmidigital.com/nxtbooks/ibmsystemsmag/ibmsystems_power_200909/index.php#/16
(**) Download http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS3780

© 2010 / 2011 IBM Corporation

SSD Analyzer Tool for IBM i

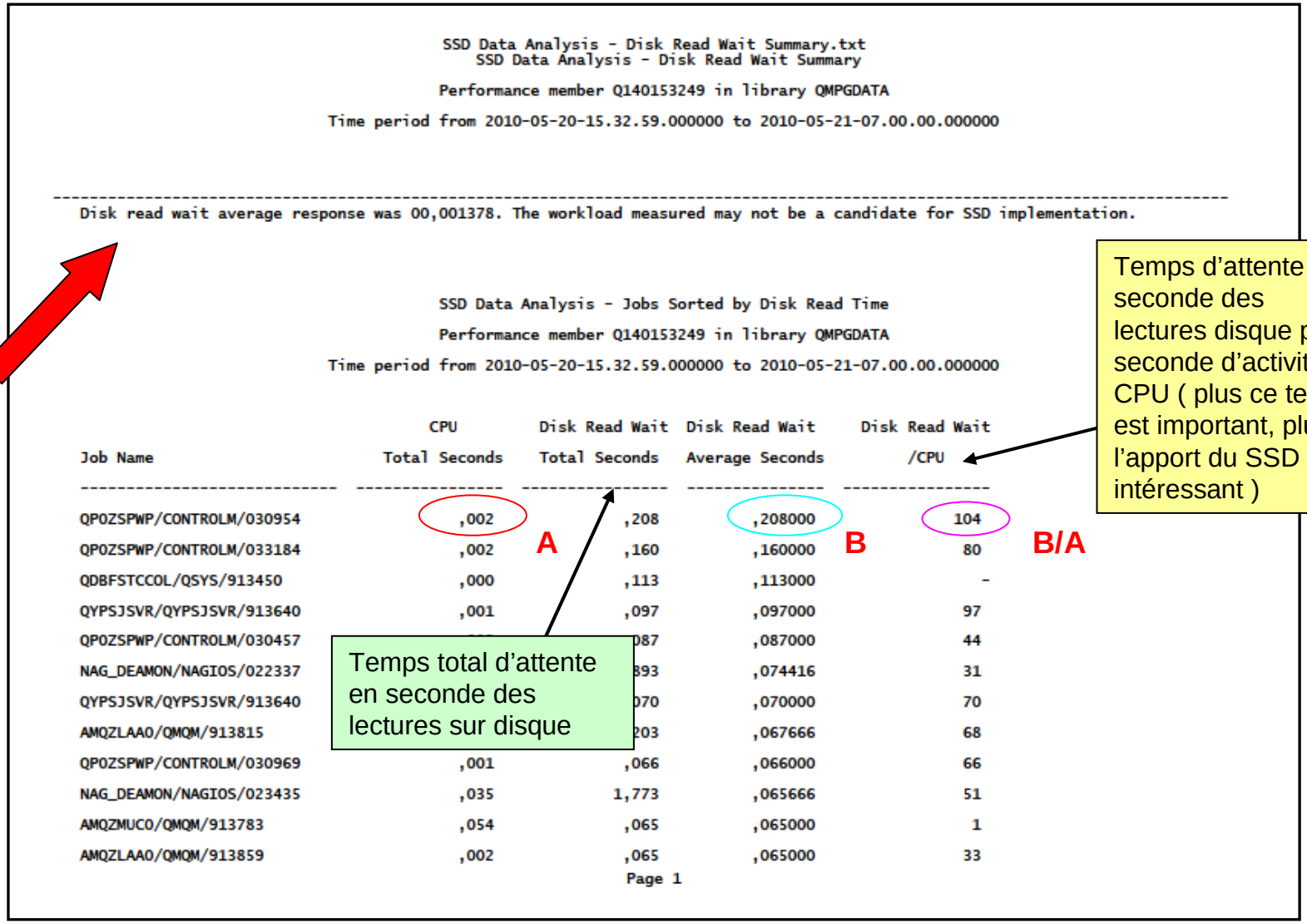
- Rapide, simple, outil d'analyse gratuit disponible en V5R4 et plus
- Se présente comme un rapport standard de performance
- Donne une indication de type "probablement oui", "probablement non", ou "peut être »
- **A venir**: estimation du nombre de SSD recommandé *(4 Q 2010 / 1 H 2011)*


SSD ANALYZER TOOL (ANZSSDDTA)
Type choices, press Enter.
PERFORMANCE MEMBER . . . . . *DEFAULT Name, *DEFAULT
LIBRARY . . . . . Name
Additional Parameters
REPORT TYPE . . . . . *SUMMARY *DETAIL,*SUMMARY,*BOTH
TIME PERIOD::
START TIME AND DATE::
BEGINNING TIME . . . . . *AVAIL Time, *AVAIL
BEGINNING DATE . . . . . *BEGIN Date, *BEGIN
END TIME AND DATE::
ENDING TIME . . . . . *AVAIL Time, *AVAIL
ENDING DATE . . . . . *END Date, *END
DETAIL REPORT SORT ORDER . . . . *DISKTOT *JOBNAME, *CPUTOT...
NUMBER OF RECORDS IN REPORT . . 50 0 - 9999, *ALL
Bottom
F3=Exit F4=Prompt F5=Refresh F12=Cancel F13=How to use this display
F24=More keys

Outil disponible sous forme de SAVF sur www.ibm.com/support/techdocs dans "Presentations & Tools". Faire une recherche sur le mot clef SSD. Attention à suivre le mode d'emploi, surtout en V5R4 *(QIBMSSD/SETUP)*

© 2010 / 2011 IBM Corporation

Conclusions de l'outil …



SSD Data Analysis - Disk Read Wait Summary.txt
SSD Data Analysis - Disk Read Wait Summary

Performance member Q140153249 in library QMPGDATA

Time period from 2010-05-20-15.32.59.000000 to 2010-05-21-07.00.00.000000

Disk read wait average response was 00,001378. The workload measured may not be a candidate for SSD implementation.

SSD Data Analysis - Jobs Sorted by Disk Read Time

Performance member Q140153249 in library QMPGDATA

Time period from 2010-05-20-15.32.59.000000 to 2010-05-21-07.00.00.000000

Job Name	CPU Total Seconds	Disk Read Wait Total Seconds	Disk Read Wait Average Seconds	Disk Read Wait /CPU
QP0ZSPWP/CONTROLM/030954	,002	,208	,208000	104
QP0ZSPWP/CONTROLM/033184	,002	,160	,160000	80
QDBFSTCCOL/QSYS/913450	,000	,113	,113000	-
QYPSJSVR/QYPSJSVR/913640	,001	,097	,097000	97
QP0ZSPWP/CONTROLM/030457		087	,087000	44
NAG_DEAMON/NAGIOS/022337		893	,074416	31
QYPSJSVR/QYPSJSVR/913640		070	,070000	70
AMQZLAA0/QMQM/913815		203	,067666	68
QP0ZSPWP/CONTROLM/030969	,001	,066	,066000	66
NAG_DEAMON/NAGIOS/023435	,035	1,773	,065666	51
AMQZMUC0/QMQM/913783	,054	,065	,065000	1
AMQZLAA0/QMQM/913859	,002	,065	,065000	33

Page 1

A

B

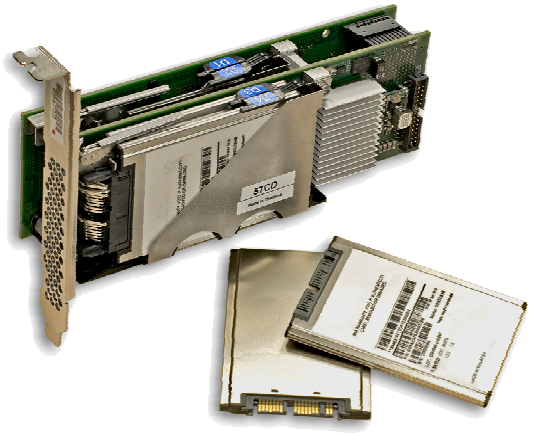
B/A

Temps d'attente en seconde des lectures disque par seconde d'activité CPU (plus ce temps est important, plus l'apport du SSD sera intéressant)

Temps total d'attente en seconde des lectures sur disque

© 2010 / 2011 IBM Corporation

Nouveautés Aout 2010



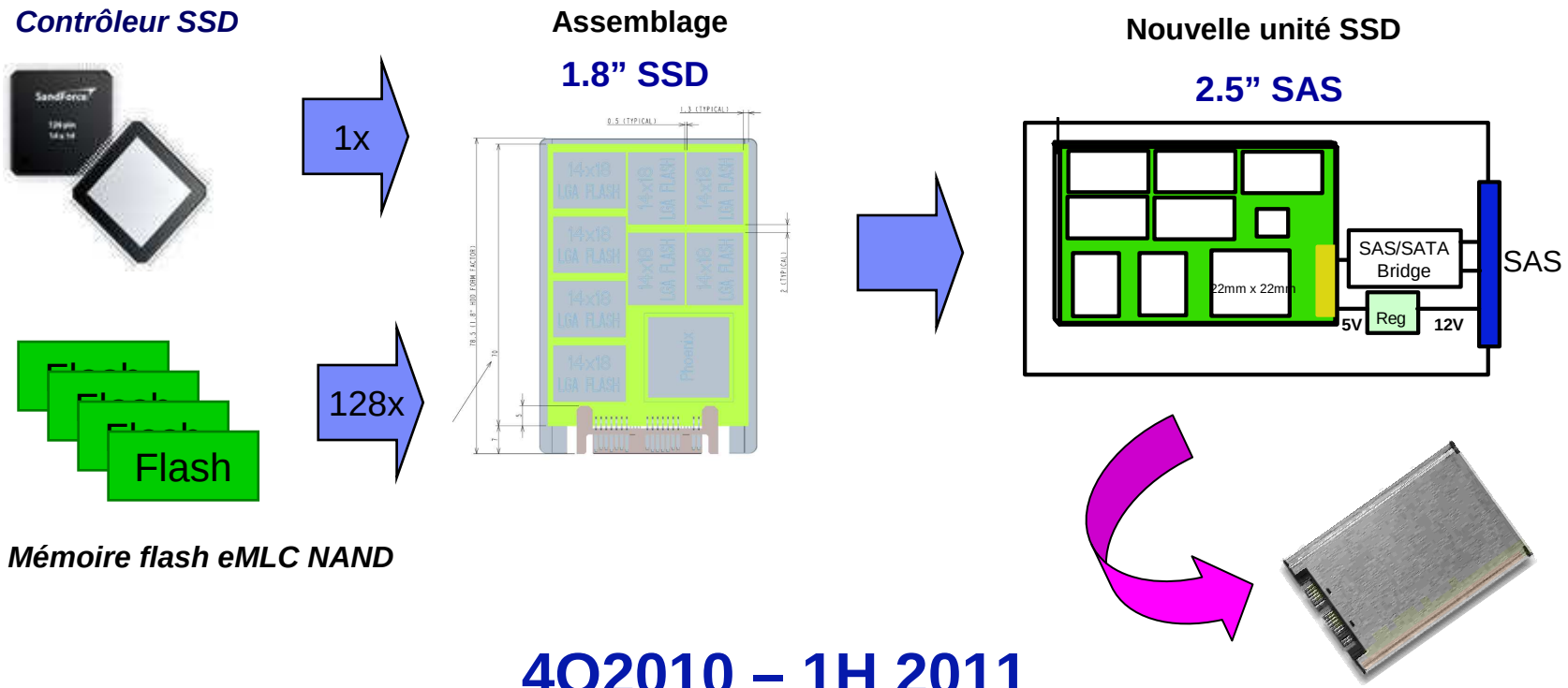
Technologie Entreprise Multi-Level Cell

© 2010 / 2011 IBM Corporation

Performances à la hausse et coûts à la baisse !!!

- **Nouvelle unité SSD 177GB - 2.5” (eMLC) pour slots HDD**
- **Nouveau format SSD 1.8” pour intégration dans cartes PCIe**

2eme génération des Flash NAND



4Q2010 – 1H 2011

© 2010 / 2011 IBM Corporation

Caractéristiques des nouvelles cartes « embeded » RAID SSD en PCIe

Dispositifs Bus Serveur

- PCIe1.0 x8 pour vitesse de transfert “up to 2Go/s”
- Format carte double et longue

Dispositifs Bus Unité

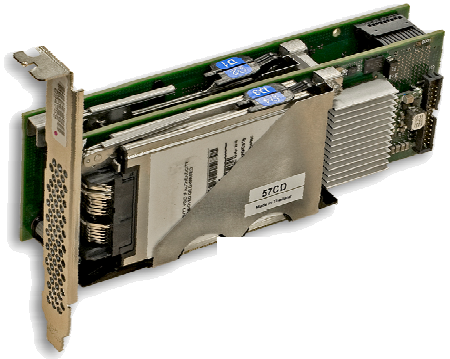
- vitesse interface SAS = 3Gb/s
- CRC 4 bytes pour chaque LBA

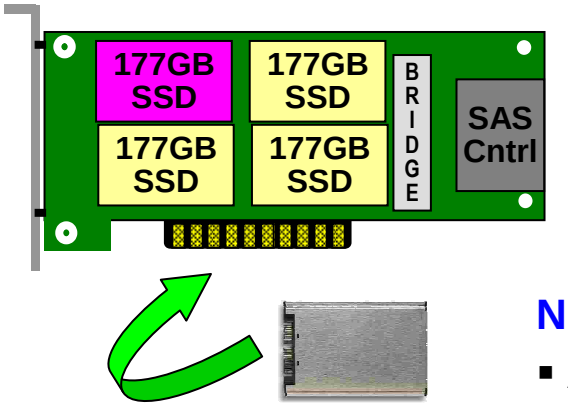
Dispositifs Mémoire

- 512Mo partagés entre C/S & D/S
- Pas de cache écriture

Performance

- Seq Rd .. 800 Mo/sec JBOD/RAID
- Seq Write .. 800 Mo/ sec JBOD
- Seq Write .. 200Mo/sec RAID 5
- **48.000 E/S par sec** en Read/Write JBOD
- **32.000 E/S par sec** en Read / R5





Compatibilité Software

- **IBM i 7.1**
- **AIX 5.3**
- **Linux … SLES 10, RHEL 5.5**

Fonctions Avancées

- de 1, 2 ou 4 sous-ensembles de 177 Go par carte
- support RAID 0,10,5
- “Hot Spare” pour RAID 5
- Contrôle de parité en arrière plan
- Auto Config en RAID 0

Unité SSD utilisée

- eMLC en 177GB utilisables
- 708Go installés (*4x177Go)*

Niveau de Service et Support

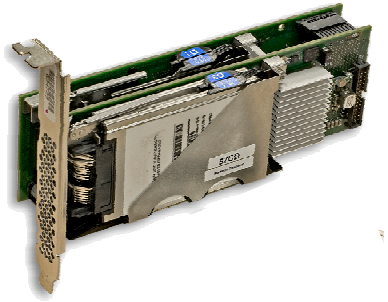
- Amélioration du RAS
- **du serveur 710 au 780**

Le stockage interne est très bien positionné pour tirer partie de la bande passante et de la faible latence de la technologie SSD

© 2010 / 2011 IBM Corporation

PCIe-SSD-based Protection

Point clef: la carte PCIe peut être enlevée du slot PCI pour remplacer un module SSD … attention: SSD “hot plug” seulement si l’adaptateur PCI et ses modules sont en configuration miroir (et sont dans un slot PCIe acceptant le « hot swap »)



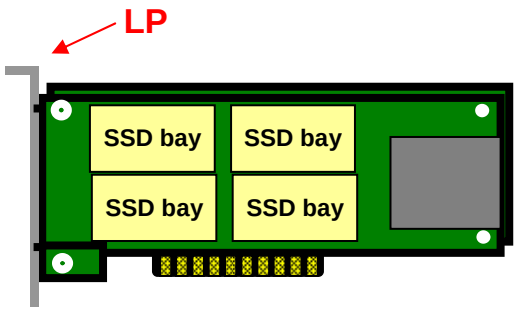
Les modules SSD sont extrêmement fiables, mais la protection contre les pannes d’unités (au même titre que les unités de disque) est FORTEMENT recommandée

- **Première option de protection**: Miroir via le système d’exploitation*
  - Adaptateurs redondants PLUS modules SSD redondants
    - » possibilité de hot plug … à choisir de préférence dans la plupart des situations
  - Capacité des SSD limitée à 50% pour la protection en miroir
- **Deuxième option de protection** : RAID 5
  - L’adaptateur n’est pas redondant .. Pas de possibilité de modules SSD « hot plug »
  - La capacité affectée à la protection des SSD est limitée à 25%
  - Option Alternative: Ajout d’une unité « hot spare » (50% de capacité pour la protection)
- **Troisième option de protection** : RAID 6
  - L’adaptateur n’est pas redondant .. Pas de possibilité de modules SSD « hot plug »
  - 50% de la capacité des SSD affectée à la protection

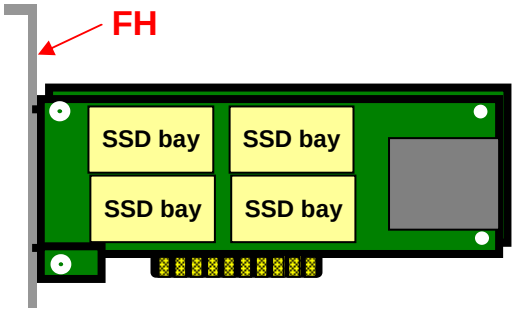
* La fonction miroir sous AIX/Linux: est disponible via le système d’exploitation aussi bien sur les fichiers JBOD ou RAID 0. La fonction miroir en IBM i n’utilise pas le RAID 0 ou le JBOD. IBM i peut utiliser indifféremment le miroir , le RAID 5 ou le RAID 6.

© 2010 / 2011 IBM Corporation

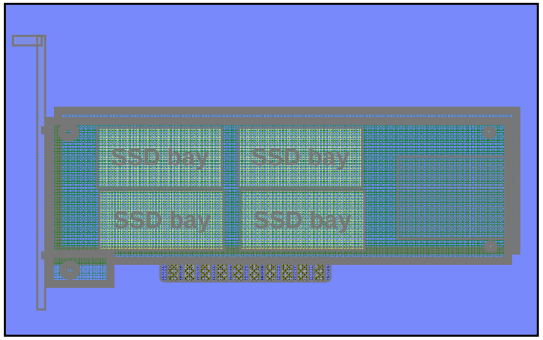
Codes dispositif pour adaptateur PCIe RAID & SSD SAS



#2053 Low Profile (LP)
… pour CEC Power
710, 720, 730, 740



#2054 Full Height (FH)
… pour CEC Power
720, 740, 750, 755



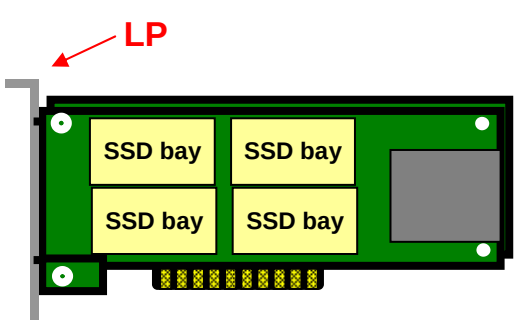
#2055 en carte double Gen3 BSC pour tiroirs E/S PCIe 12X (#5802/5877) sur 720/740/750/770/780

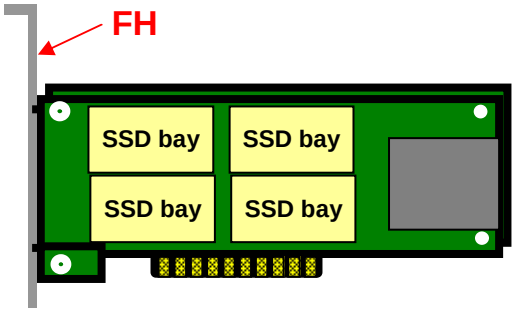
Cartes non supportées sur CEC 770/780, mais dans tiroir E/S.

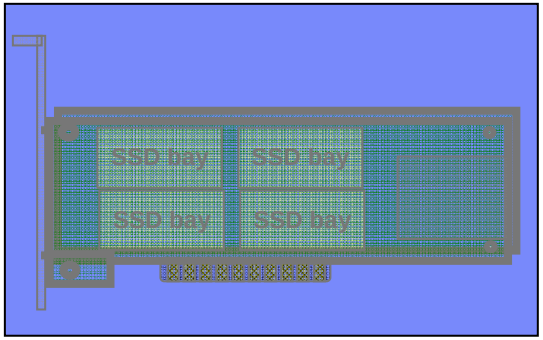
- **Trois codes dispositif possibles**
- **Electronique identique – même CCIN (57CD)**
- **Les dispositifs #2053 et #2054 ont des hauteurs d'équerre différentes pour s'adapter aux slots PCIe de type « Low Profile » ou standard**
- **Le dispositif #2055 est un #2054 dans une BSC (Blind Swap Cassette) double épaisseur**
- **Les adaptateurs doivent être commandés avec, au moins, un module SSD**

© 2010 / 2011 IBM Corporation

Codes dispositif pour adaptateur PCIe RAID & SSD SAS







#2053 Low Profile (LP)

#2055 en carte double

carte #2053 LP temporairement non supportée sur 720/740 en tour

carte #2054 FH temporairement suspendue suite à un problème technique lié à l'emplacement de la carte. Solution attendue en 2011

Pas de problème pour cette carte #2055

- **Le dispositif #2055 est u** **l'équerre différentes pour** **u standard** **lind Swap Cassette)** **double épaisseur**
- **Les adaptateurs doivent** **u moins, un module SSD**

© 2010 / 2011 IBM Corporation

Codes dispositif pour module SSD 177GB en technologie eMLC

#1995 Module 177GB pour AIX et Linux

#1996 Module 177GB pour IBM i

- **Deux codes dispositif possibles**
- **Physiquement et électroniquement identiques et interchangeables – même code CCIN (58B2)**
- **Deux codes dispositif utilisés par les configurateurs d'IBM (eConfig/SPT) pour permettre à l'usine d'identifier l'usage du SSD (*les règles de protection variant en fonction de l'usage des serveurs)***

© 2010 / 2011 IBM Corporation

Que choisir: SSD-SAS en tiroir ou carte PCIe ??

En 2010, les modules eMLC sur carte PCIe SAS présentent deux fois plus de capacité que les SSD SLC traditionnels, MAIS attention à bien analyser votre besoin (niveau de version, technologie des serveurs Power, protection souhaitée …)

SSD sur carte PCIe …

- Miroir: utilise 4 slots PCIe pour une capacité jusqu'à 708 Go
- RAID-5: utilise 2 slots PCIe pour une capacité de 531 Go, 304 Go si "hot spare"
- RAID-6: utilise 2 slots PCIe pour une capacité de 304 Go

SSD sur attachement SAS …

- Supporte le "Hot Plug" si baie SAS et contrôleur SAS séparés
- Paire d'adaptateurs SAS #5903, utilise 2 slots PCIe; jusqu'à 9 SSD
  - En protection RAID 5 = 557 Go dans un tiroir E/S 12X IB PCIe #5802
- Adaptateur #5904/5906/5908, utilise 2 slots PCI-X; jusqu'à 8 SSD
  - En protection RAID 5 = 388 Go dans une tiroir disque EXP12S #5886

© 2010 / 2011 IBM Corporation

Détails sur le support des cartes PCIe

Système d'exploitation ….

- AIX Version 7.1
- AIX Version 6.1 avec 6100-06 Technology Level
- AIX Version 5.3 avec 5300-12 Technology Level et Service Pack 2, ou plus (disponibilité prévue : 30 Sept 2010)
- AIX Version 5.3 avec 5300-11 Technology Level et Service Pack 5, ou plus (disponibilité prévue : 30 Sept 2010)
- AIX Version 5.3 avec 5300-10 Technology Level et Service Pack 5, ou plus (disponibilité prévue : 30 Sept 2010)

- IBM i 7.1 ou plus

- SuSE Linux Enterprise Server Version 10, Service Pack 3 ou plus
- SuSE Linux Enterprise Server Version 11, Service Pack 1 ou plus
- Red Hat Enterprise Linux Version 5.5 ou plus

- VIOS 2.2 minimum.

© 2010 / 2011 IBM Corporation

En complément …

- Le nouvel adaptateur PCIe, de même que la solution SSD sur baie SAS, supporte la fonction d'IPL, en mode "load source" comme en mode "alternate IPL"

- Les SSD peuvent être utilisés en lieu et place des disques traditionnels dans le CEC. Ils peuvent même être une alternative (onéreuse !!) aux disques classiques sur la partition/serveur.

- Ce nouvel adaptateur PCIe, comme la solution SSD sur baie SAS, peut être utilisé par PowerVM (VIOS)

- **Fonction MIROIR:**
  - De même que les SSD classiques SAS 69 Go, les SSD 177 Go sur PCIe ne peuvent pas être en configuration miroir sur un HDD classique.
  - Le SSD sur adaptateur PCIe doit être en miroir sur un SSD de même taille lui même sur adaptateur PCIe.
  - Un SSD de 177 Go sur adaptateur PCIe ne peut pas être mis en miroir sur un SSD de 69 Go en baie SAS.

© 2010 / 2011 IBM Corporation

Glossaire …

- SCSI ... Small Computer System Interface (*technologie en fin de vie)*
- SAS … Serial Attached SCSI
- HDD …Hard Disk Drive
- SSD… Solid State Drive
- SAN … Storage Area Network
- NPIV … N_Port ID Virtualization
- VIOS … Virtual I/O Server
- SFF … Small Form Factor
- IOA … Input/Output Adapter
- IOP … Input/Output Processor
- Smart IOA … Intelligent I/O Adapter
- PCI-x … PCI eXtended (enhanced PCI card and slot)
- PCIe … PCI Express (latest and fastest enhanced PCI card and slot)
- HSL … High Speed Loop (POWER4 thru POWER6 I/O bus interconnect)
- RIO … Remote I/O -same as HSL, but called RIO when used on p system
- 12X … IBM's POWER System implementation of InfiniBand bus interconnect
- CEC … Central Electronics Complex *- refers to the processor enclosure for POWER Systems*

© 2010 / 2011 IBM Corporation

Documentations

© 2010 / 2011 IBM Corporation

AIX SSD configuration

- **Devices are initially configured as pdisks**

```
# lsdev -Cc pdisk
pdisk0 Available 02-08-00 Physical SAS Disk Drive
pdisk1 Available 02-08-00 Physical SAS Disk Drive
```

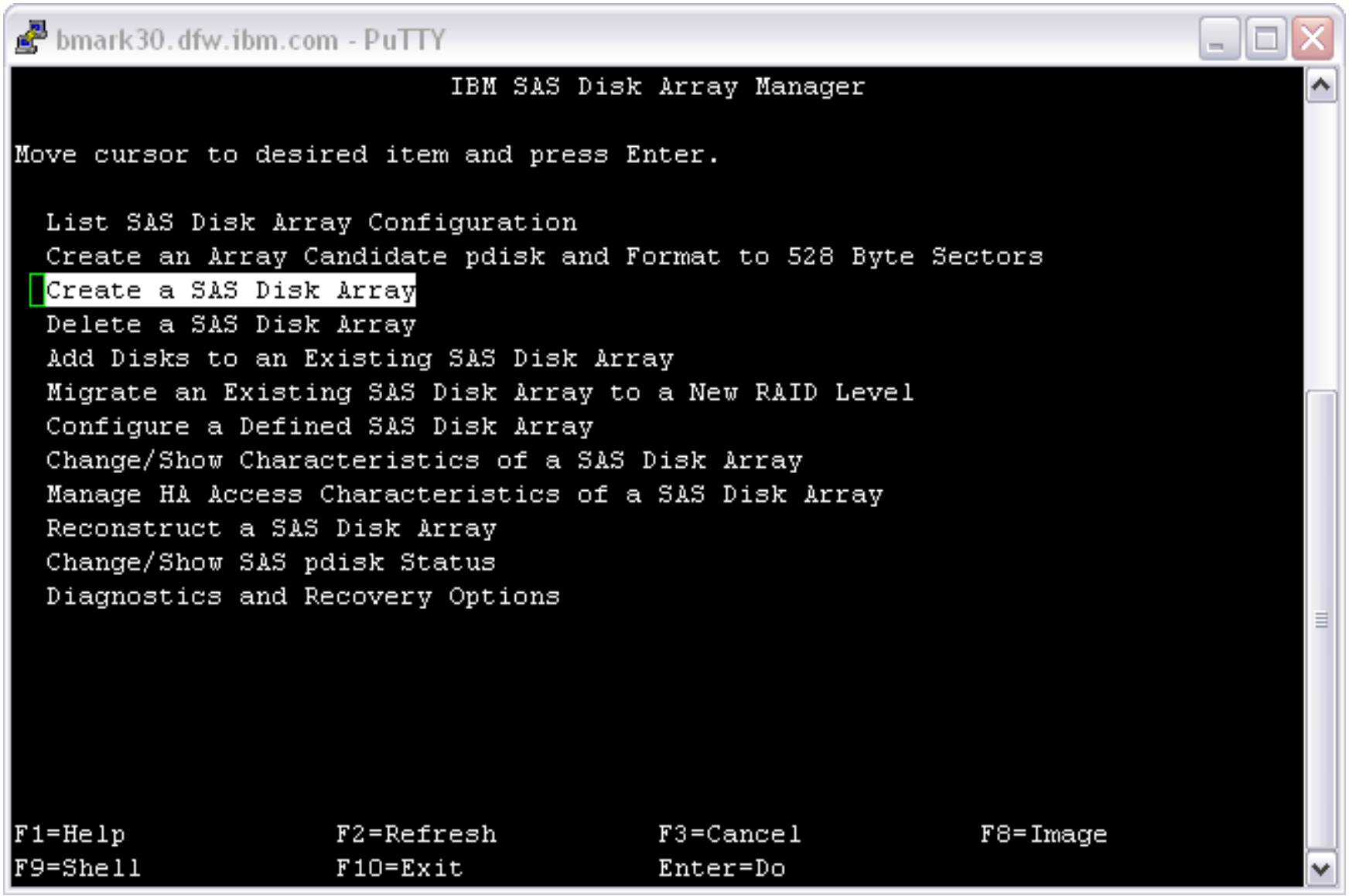
- **Smitty devices -> Disk Array -> IBM SAS Disk Array -> IBM SAS Disk Array Manager**
- **Create an Array Candidate pdisk and Format to 528 Byte Sectors**
- **Create a SAS RAID array**
  - ✓ RAID 0, 5, 6 or 10 (2 disk RAID 10 = RAID 1)
  - ✓ 16, 64 or 256 KB stripe (aka strip) size
    - ➢ RAID 5 will be popular
  - ✓ An hdisk appears

```
# lsdev -Cc disk | grep "SAS RAID"
hdisk3 Available 02-08-00 SAS RAID 0 Disk Array
hdisk5 Available 02-08-00 SAS RAID 0 Disk Array
```

- **Choose whether or not to turn on write cache for the adapter**
- **Probably do not create hot spares**
- **Proceed to LVM configuration**

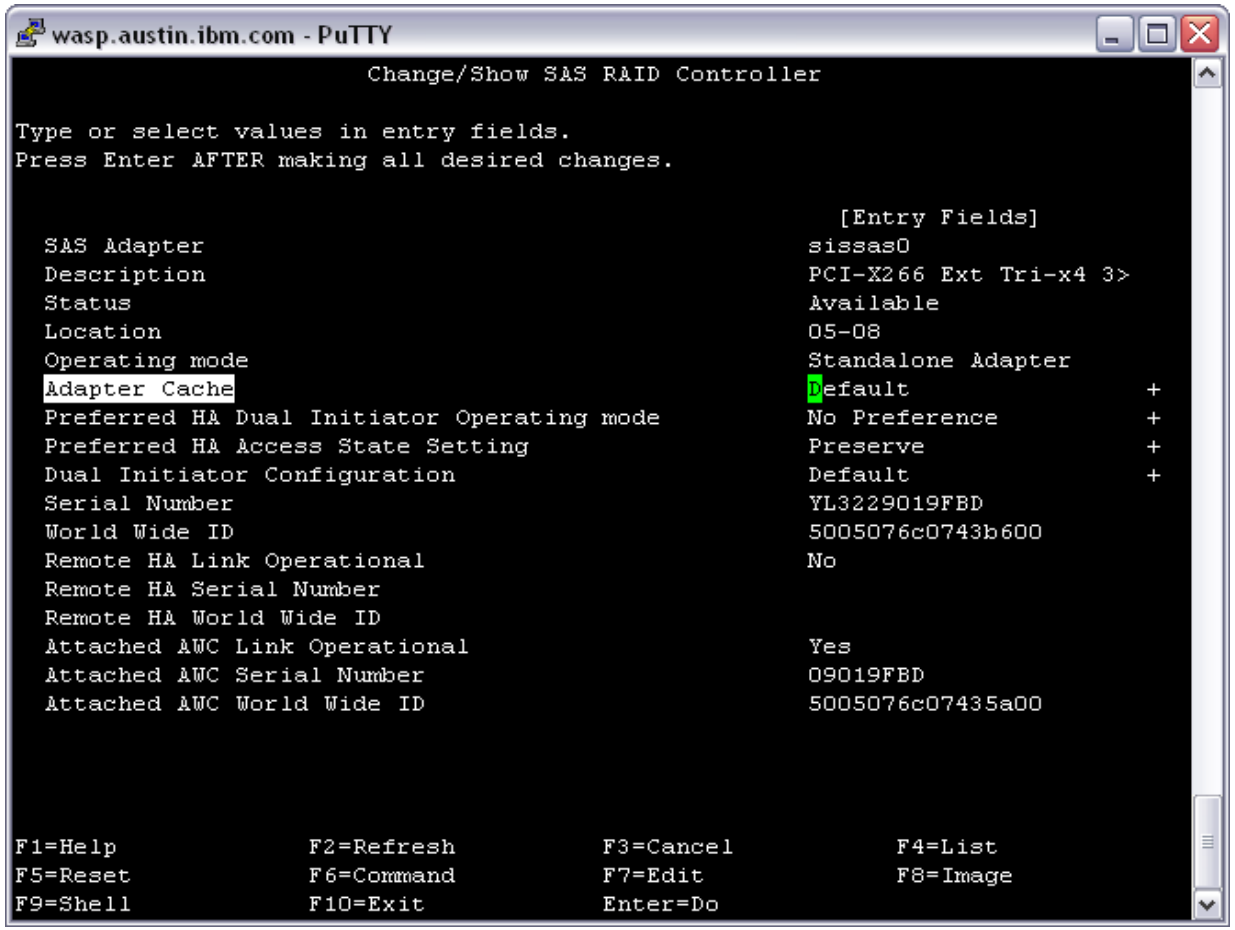
© 2010 / 2011 IBM Corporation

AIX SSD configuration


bmark30.dfw.ibm.com - PuTTY
IBM SAS Disk Array Manager
Move cursor to desired item and press Enter.
List SAS Disk Array Configuration
Create an Array Candidate pdisk and Format to 528 Byte Sectors
Create a SAS Disk Array
Delete a SAS Disk Array
Add Disks to an Existing SAS Disk Array
Migrate an Existing SAS Disk Array to a New RAID Level
Configure a Defined SAS Disk Array
Change/Show Characteristics of a SAS Disk Array
Manage HA Access Characteristics of a SAS Disk Array
Reconstruct a SAS Disk Array
Change/Show SAS pdisk Status
Diagnostics and Recovery Options
F1=Help
F2=Refresh
F3=Cancel
F8=Image
F9=Shell
F10=Exit
Enter=Do

© 2010 / 2011 IBM Corporation

AIX SSD configuration


wasp.austin.ibm.com - PuTTY
Change/Show SAS RAID Controller
Type or select values in entry fields.
Press Enter AFTER making all desired changes.
[Entry Fields]
SAS Adapter sissas0
Description PCI-X266 Ext Tri-x4 3>
Status Available
Location 05-08
Operating mode Standalone Adapter
Adapter Cache Default +
Preferred HA Dual Initiator Operating mode No Preference +
Preferred HA Access State Setting Preserve +
Dual Initiator Configuration Default +
Serial Number YL3229019FBD
World Wide ID 5005076c0743b600
Remote HA Link Operational No
Remote HA Serial Number
Remote HA World Wide ID
Attached AWC Link Operational Yes
Attached AWC Serial Number 09019FBD
Attached AWC World Wide ID 5005076c07435a00
F1=Help F2=Refresh F3=Cancel F4=List
F5=Reset F6=Command F7=Edit F8=Image
F9=Shell F10=Exit Enter=Do

- **From Disk Array Manager menu -> Diagnostics and Recovery Options -> Change/Show SAS RAID Controller**
- **Adapter Cache can be set to Disabled**
  - ***For adapters in HA configurations with SSDs, disabled cache probably is best***

© 2010 / 2011 IBM Corporation

Choosing data to place on SSDs

- ***iostat* – identify IOPS (tps) and R/W ratio for PV**
  - ✓ High small-block tps and high R/W ratio suggests good candidate
  - ✓ *iostat –t*: if there is no iowait time, SSDs will not improve performance
  - ✓ *iostat –D*:
    Total system storage *tps* may be substantially higher SDD vs HDD
    look for hdisks that do over 200 IOPS (*tps*) or *%tm act* at least 99% busy
  - ✓ Investigate further with *lvmstat* and # lspv –l <hdisk#>
- **lvmstat – identify IOPS (iocnt) and R/W ratio for LVs**
  - ✓ Turn on *lvmstat* for VG with # lvmstat –e –v <vgname>

```
# lvmstat -v newvg2
Logical Volume         iocnt   Kb_read   Kb_wrtn      Kbps
…
```

  - ✓ High iocnt and high R/W ratio LVs are good candidates
  - ✓ Also reports IOPS on PPs – useful when the LV is relatively large
- ***filemon* – identify IO sizes, sequentiality to PVs, LVs**
  * trace for only seconds of time
  * see sorted LV utilization in “Most Active Logical Volumes” report. High utilization is good SSD candidate.
- **Some applications have tools for identifying hot data**
  - ✓ DB2 snapshot monitoring tool

© 2010 / 2011 IBM Corporation

Moving data to SSDs

- Add SSD hdisk to VG with

```
# extendvg <vgname> <hdisk#>
```

- Migrate LV to hdisk dynamically with

```
# migratepv –l <lvname> <source hdisk> <target hdisk(s)>
```

  - Repeat with other hdisks the LV resides on

- Or create a new VG for just SSD data

```
# mkvg –y ssdvg –s 32 –S <ssdhdisk(s)>
```

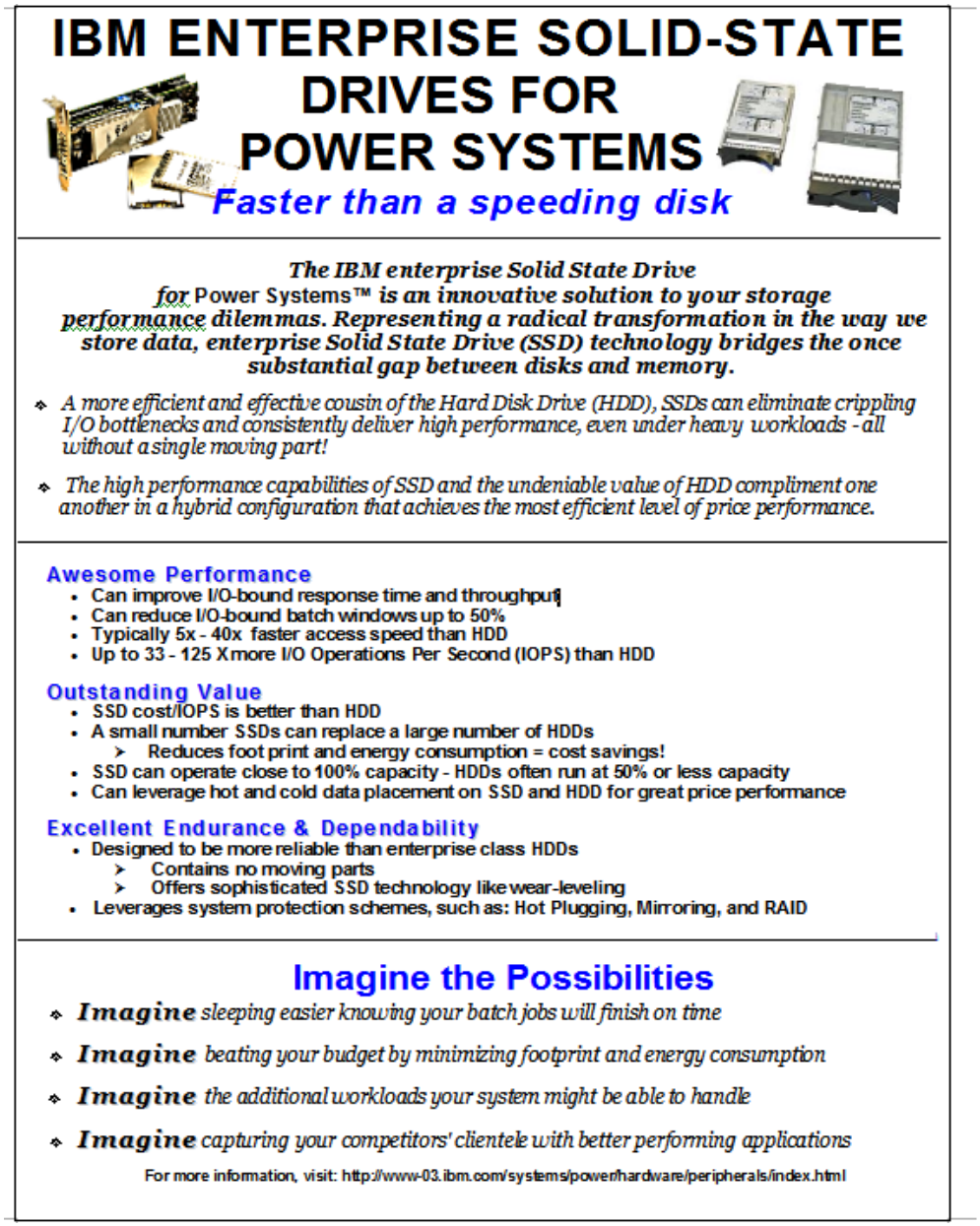
  - Offers smaller PP sizes to waste less space

- Stop application and copy the LV to the new VG with
  - `# cplv –v <sourcevg> -y ssdlv <source LV>`
  - Or backup/restore/copy data to new file system

© 2010 / 2011 IBM Corporation

Brochure sur SSD

Accès sur pages web SSI / PW



IBM ENTERPRISE SOLID-STATE DRIVES FOR POWER SYSTEMS

Faster than a speeding disk

The IBM enterprise Solid State Drive for Power Systems™ is an innovative solution to your storage performance dilemmas. Representing a radical transformation in the way we store data, enterprise Solid State Drive (SSD) technology bridges the once substantial gap between disks and memory.

- *A more efficient and effective cousin of the Hard Disk Drive (HDD), SSDs can eliminate crippling I/O bottlenecks and consistently deliver high performance, even under heavy workloads - all without a single moving part!*
- *The high performance capabilities of SSD and the undeniable value of HDD compliment one another in a hybrid configuration that achieves the most efficient level of price performance.*

Awesome Performance

- Can improve I/O-bound response time and throughput
- Can reduce I/O-bound batch windows up to 50%
- Typically 5x - 40x faster access speed than HDD
- Up to 33 - 125 X more I/O Operations Per Second (IOPS) than HDD

Outstanding Value

- SSD cost/IOPS is better than HDD
- A small number SSDs can replace a large number of HDDs
  - Reduces foot print and energy consumption = cost savings!
- SSD can operate close to 100% capacity - HDDs often run at 50% or less capacity
- Can leverage hot and cold data placement on SSD and HDD for great price performance

Excellent Endurance & Dependability

- Designed to be more reliable than enterprise class HDDs
  - Contains no moving parts
  - Offers sophisticated SSD technology like wear-leveling
- Leverages system protection schemes, such as: Hot Plugging, Mirroring, and RAID

Imagine the Possibilities

- ***Imagine*** *sleeping easier knowing your batch jobs will finish on time*
- ***Imagine*** *beating your budget by minimizing footprint and energy consumption*
- ***Imagine*** *the additional workloads your system might be able to handle*
- ***Imagine*** *capturing your competitors' clientele with better performing applications*

For more information, visit: http://www-03.ibm.com/systems/power/hardware/peripherals/index.html

© 2010 / 2011 IBM Corporation

Cinq “White Papers” sur SSD en Power Systems publiés en 2009

IBM Power SSD vs Consumer SSD **(publié en Nov 2009)**
- “Advantages of True Enterprise Solid State Drives (SSDs) in Enterprise Systems”

AIX-specific **(publié en Avril 2009)**
- “Driving Business Value on Power Systems with Solid State Drives”

IBM i-specific **(publié en Mai 2009)**
- “Performance Value of Solid State Drives using IBM i”
- First published May 2009

Plus spécifiquement sur la technologie SSD – fonction d’AIX/IBM i/Linux **(publié en Juin 2009)**
- “Performance Impacts of Flash SSDs Upon IBM Power Systems”

Ces documents sont disponibles sur le site des Power Systems dans la rubrique “Resources/Literature”
- http://www.ibm.com/common/ssi/apilite?infotype=SA&infosubt=WH&lastdays=1825&hitlimit=200&ctvwcode=US&pubno=POW*USEN&appname=STGE_PO_PO_USEN_WH&additional=summary&contents=keeponlit

5eme document sur l’environnement SAP
- http://www.sdn.sap.com/irj/sdn/db4?rid=/library/uuid/90a1637e-065f-2c10-3ab7-bea9375fc88d

© 2010 / 2011 IBM Corporation

Références/Outil pour IBM i et AIX

IBM - Performance Management on IBM i Resource Library
http://www.ibm.com/systems/i/advantages/perfmgmt/resource.html

Performance Value of Solid State Drives using IBM i
http://www.ibm.com/systems/resources/ssd_ibmi.pdf

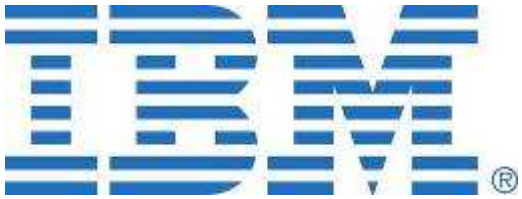
Performance Impacts of Flash SSDs Upon IBM Power Systems
http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=POW03028USEN&attachment=POW03028USEN.PDF&appname=STGE_PO_PO_USEN_WH

Driving Business Value on Power Systems with Solid State Drives
http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=POW03025USEN&attachment=POW03025USEN.PDF&appname=STGE_PO_PO_USEN_WH

IBM Systems Lab Services and Training
http://www.ibm.com/systems/services/labservices

IBM Power Systems(i) Benchmarking and Proof-of-Concept Centers
http://www.ibm.com/systems/i/support/benchmarkcenters

© 2010 / 2011 IBM Corporation



© 2010 / 2011 IBM Corporation